21/07/2021

Master Thesis

Business process optimization: An approach for improving organizations by integration of external data

Nick Kerckhoffs
Master Student Business Information Technology

Supervisors:

Dr. M.Daneva
University of Twente

Dr.ir. H.Moonen
University of Twente

Simon Doesburg
Lumen BS

Daan Witjes
Van Eijck

Preface

This master thesis is the last step in completing my master business information technology at the University of Twente. The last six years studying in Enschede have been a joy and I look forward to what is about to come in the next years.

Firstly, I would like to thank Lumen BS for facilitating me with this research. The people within Lumen were always supportive during my research and were always available when I had questions about the tools I had to use or other questions in general. I would like to thank Simon Doesburg especially for being my supervisor from Lumen. Thanks for the helpful insights during our meetings and general interest in me and this project.

Secondly, I would like to thank Van Eijck for providing me with the environment to perform my research. From Van Eijck I would like to thank Daan Witjes in particular who was always directly available to teach me the ways of the salvage sector from which I first knew nothing and teach me all the different meanings of the data field of their BI system.

Thirdly, I would like to thank my supervisors from the University of Twente, Maya Daneva and Hans Moonen, for helping me shape this thesis to its current form with all their feedback, insightful meeting and suggestions.

Lastly, I would like to thank my roommates and girlfriend for the support during my master thesis. Because of the corona quarantine I have seen my roommates even more often than usual during the week and they have always been supportive and my girlfriend was also always there to provide support or even feedback during my master thesis project.

Abstract

The use of external data sources becomes more popular every day. The amount of data that is available keeps on growing and services like linked open data become more usable. Organizations generate a lot of data themselves with the use of CRM-systems. These systems have as goal to create a better customer experience, which will hopefully improve sales. The goal of this research was to create an approach that organizations can use to prepare themselves for the integration of external data, implement it, analyze the connections between the internal and external data and make changes based on the analysis. This approach has been created using a literature research, interviews with experts and empirical data from a case study. The literature research looked at eight papers about CRM-system data analysis models and ranked these papers based on the strength and weaknesses of these models. The strengths of these models, namely the combined focus of models on the business and the technical side have been used for the creation of the final approach.

During this research a case study has been held at Van Eijck, a salvage company in the Netherlands. The goal of this case study was to integrate external information in the database of Van Eijck and improve business processes in their organization based on findings in the internal/external data analysis. Business processes could be improved by lowering the time it takes for employees of Van Eijck to arrive at an incident. The data that was chosen as external data was weather data. Through the analysis of the data of Van Eijck and the weather data, numerous visualizations have been built and multiple findings were made. These findings were used to advise Van Eijck on how they could improve their business. They can do this by focusing on the specific rayons which are most influenced by the weather and keep a close eye on the performance dashboard of their rayons that has been made during this research. The empirical data from this case study combined with data gathered during interviews with stakeholders has been used to create the final approach.

The end product of this research, the approach for external data integration and analysis has been created using the steps above consists of the following steps: a preparation of the business side, a preparation of the technical side, choosing the external data, choosing the source, interesting the data, finding connections, implementing changes and an evaluation. By following these steps organizations can more easily and in a structured way improve business processes with the help of external data. The approach has been evaluated by experts in the field of system analysis by using an empirically tested questionnaire and by means of a group interview. The overall conclusion of the evaluation is that the approach is usable and useful. The approach created can be used as a general guideline when wanting to make changes in your organization and can be combined with existing approaches when one of the steps in the approach is not clear enough according to the experts. Finally, this thesis provides a discussion on limitations and recommendation for the participating organizations in the research.

Table of Contents

Chapter 1. Introduction

This study was performed as part of the master program business information and technology at the University of Twente. The research was conducted with the help of two organizations, Lumen BS and Van Eijck. This first chapter will introduce the problems that this master thesis will try to solve and why these problems should be solved. The problem is divided into multiple research questions which will be discussed together with the method that will be used to answer the research questions.

1.1 Problem identification and motivation

The use of big data sets becomes more and more popular among organizations nowadays. Through means of data analysis it becomes increasingly interesting to analyze the data that an organization produces, because this analysis can produce visualization on how the organization is currently performing. Using this current analysis the organization can determine based on key performance indicators how they are doing and where they want to focus on in order to make their company more productive in the future. These current analyses are not the only thing that can be generated using data analysis, they can also produce prediction models. These models show how the company will perform in the future if they keep their processes the same. The models can also be used to see what should change in the organization in order to positively influence the prediction models.

Besides big data, linked open data has also become increasingly popular among researchers. These are big public data sets available to everyone which can be combined with your own data to generate meaningful conclusions on how external factors influence your organization. This type of data is external from the organization, which means that it is not produced by the organization and they have little to no influence on it. But the external factors behind the external data might influence the organization. This is why it is very interesting for companies to include this type of information in their database. The combination of this data and their own might lead to unique findings and improvements which would not have been found if the focus of the analysis would solely be on their own data.

1.2 Research context

The focus of this research will be on combining data from internal systems with external data. The literature reviewed in this study is about data that is produced by CRM systems, the internal data from the case study that will be held also originates from a CRM system. This is why the context of this research is around CRM-system. CRM stands for customer relationship management and is a popular tool used by companies. CRM technology applications link front office (e.g. sales, marketing and customer service) and back office (e.g. financial, operations, logistics and human resources) (Chen. I & Popovich. K, 2003). Using this technology information about customers can be stored but it can also be used to get information about the employees. Modern companies enable their employees to use online tools to know which tasks are done at what time and by who. Based on this information a database can be filled and used in data analysis. The goal of CRM is to retain and acquire more customers through the information that is provided in earlier instances in the CRM system. One of the ways to retain and increase your amount of customers is by improving processes within the organization, to make your organization more appealing for new customers. This could be an improvement of availability, the speed at which processes are performed, or an increase in overall productivity. These aspects do not only make the organization more appealing for new customers but also ensures that the organization can handle more clients.

The research of this thesis is done through a case study at a salvage company, Van Eijck group. Van Eijck group has purchased a CRM system from Lumen BS which will assist me during my research. Van Eijck group assists cars or bigger vehicles that have broken down in the south of the Netherlands

and also internationally since 2006, by performing services in Germany and in Belgium. This help can be provided by fixing the car if it is possible, but if the car is not fixable they provide the service to tow it away. In May 2020 they decided to implement a Microsoft Dynamics 365 CRM system in their company which made it possible to create better reports on how the company was performing, Power BI was introduced in October 2020 to make it easier to visualize the results from the CRM system. However, more profit can be taken out of the system than currently is happening, by acting more actively on the data from the CRM system. Therefore I will investigate their data set and collect information within the company to improve business processes. By investigating their data, aspects of the organization might be found that can be improved, based on the data. By improving these parts in the organization, Van Eijck can transform itself into a better functioning machine than it currently is. Lumen BS is the provider of the CRM system to Van Eijck group. Lumen is a partner of Microsoft and provides ERP and CRM system to clients using the cloud or on-premise. Besides providing the systems they also provide information to their clients about the systems and they make sure the systems stay up to date. For this research, the focus will be on the CRM systems that they provide. This system makes use of Microsoft's Power BI. Power BI is an interactive tool that is used to visualize data with as purpose to increase business intelligence.

1.3 Research objective

The goal of this research is to develop an approach which helps companies improve their processes by showing how they could best combine their own data sets and external data sets and find connections to improve their organization. The company that will be worked together with during this research, Lumen BS, currently only helps customers with their internal data, but it would be beneficial for them and their customers if they were able to also use external data and have an approach that shows what the most effective way is for an internal combined with external data analysis using CRM systems, they could more easily help future customers. This is why the goal is to develop this approach based on current literature, interviews with experts in the field of data analysis, and a Case study at Van Eijck. The goal at Van Eijck is to already provide them with a recommendation based on an internal/external data analysis. Using the experience of this data analysis the model will be designed and presented. To validate the approach experts at both Van Eijck and Lumen BS will be asked to evaluate this approach.

This new approach will provide value for the companies that cooperated during the case study and other organizations that are looking to enrich their database by using external data. The goal is to ease the improvement of processes using this external data. This means that the main objectives of this research are (1) to develop an approach on how external and internal data can be combined and analyzed to improve processes within an organization (2) validate this model with the help of experts in the field of data analysis (3) provide the investigated company with useful advice based on the model to improve their processes.

The scope of my research will be companies that make use of CRM systems and more specifically for the results of the case, companies that are active in the salvage sector or other service providing companies. The results of my thesis can be used to improve companies' processes within this scope. Companies outside the service sector that make use of CRM systems can make use of the results, for them the approach developed in this research will be relevant, but the recommended improvement techniques for process optimization will not be relevant due to being specifically tailored for the service sector. The results from this research might also apply to organizations that get their data from different sources than CRM-systems, but since the current case and literature of this research focus on CRM-systems, it cannot be said for sure.

1.4 Research Question

From the objectives of the previous section, the following research questions have been formulated. To answer these research questions, existing literature is used, interviews are held and a method is proposed to achieve the objective.

The main research question is:

What is an appropriate approach that organization can use to perform a tool based analysis to improve processes using external data integration based on current literature, experts and a case study?

The goal of the main research question is to create based on the sub-research question an approach to analyze internal data from a CRM system combined with external data fitting to the company under study to increase process speed within the organization. This approach will look how businesses can prepare for external data, how they should choose it and how they should implement it. For this answer it is important to understand how current CRM data analysis are performed and what ways are already used to analyze a database with internal and external data. Once this is known it is also important to understand what tools can be used to improve the organization under study based on the findings and validate if the used model is indeed optimal for process optimization.

This question will be answered with the help of a series of sub-questions. These are the following:

RQ1. What is the state of art of process optimization using system-generated data?

The goal of this question is to learn about process optimization using system-generated data from a literature review. The literature review will look at models for process optimization and using this information the state of art of process optimization will be determined. This information will be used for the creation of the first steps of the approach for external data integration that will be created to answer the main research question.

RQ2. What are the elements which should be focussed on when improving process optimization using system-generated data?

The goal of this research question is to find the best elements from each model that are inspected in the literature review. By finding the strongest elements from these models for process optimization, the final approach for process optimization using external data integration can be enriched with the found elements.

RQ3. What is the current best-practice in industry to find connections between internal and external data?

The goal of this question is to acquire more knowledge for the analysis part of this research. To perform the research there must be an understanding of current analysis techniques for internal and external data. This will be done through interviews conducted with experts in the field of data analysis. Their knowledge will be the starting point for the analysis of this research.

RQ4. What goals do the stakeholders of the Van Eijck case of optimizing processes using weather data have, which is used for the development and evaluation of the external data integration approach?

The goal of this question is to find out what drives the stakeholders which are associated with this project. The approach which will be created for the main research questions will be made with the

help of a case study at Van Eijck. It is important to know what the goal is of Van Eijck and other stakeholders to give context to the case study and get the most out of the analysis, to create the external data integration approach.

RQ5. How can we integrate external data and analyse it based on the case study in such a way that it takes less time and effort than is currently needed for the integration and analysis?

The goal of this question is to find out based on the experiences of the Van Eijck case what the best practice is for integration of external data and how this should be analysed such that time and effort of the integration and analysis is reduced. These findings will be used in the approach for external data integration.

RQ6. How to design an approach that fits the goals of the main research question based on current literature, experts and the case study?

The goal of this research question is to create a first version of the approach for external data integration based on the findings of the previous research questions. This approach will help organizations to integrate external data into their organization, analyse is and make changes within the organization based on the findings of the analysis to improve processes.

RQ7. How can the usability of the proposed approach be improved based on an evaluation by experts?

The goal of this research question is to evaluate whether or not the proposed way of working and its recommendations are valid and that the way of working can be teached to employees of the organization so they can use it without needing the full instruction from the master thesis, but only the approach part. The validation will be done by a panel of experts from both Van Eijck, Lumen BS and a professor to see if they agree with the tool and the solutions and can reproduce them. Their feedback will also be used to determine how the approach can be improved in future iterations.

RQ8. How can the proposed approach help organizations in general improve their processes with external data compared to existing external data integration approaches?

The goal of this research question is to find out how the created approach for external data integration can be used on only by organizations in the service sector but by organizations in general. This will be done by looking at literature and the advice from experts.

1.5 Research Methodology

For the research methodology a combination of frameworks will be used. The first framework is the Design Science Research Methodology(DSRM) (K. Peffers et al, 2007). The objective of this methodology is to develop a conceptual process for design science research in IS (information systems) and a mental model for its presentation. This means that using this framework a new concept will be developed to benefit the research domain and this concept will be explained with the help of a model to clarify the use and how to use the concept. DSRM exists out of six different phases.

The first phase of DSRM is problem identification and motivation. In this phase the specific goal for the research will be defined and the value of the solution. The state of the problem will show the importance of the solutions and thus give meaning to the research. The problem identification is covered in chapter one of this research.

The next phase is the objectives of the solution. In this phase, based on the problem that has been identified in the previous phase, the objectives of the solution are formed. The desirable solution will be described with the help of quantitative terms. This is done in chapters one and two and used to answer the first sub-question.

Design and development is the following phase. Here the actual artifact will be created, based on the desired functionalities that the artifact must have to satisfy the end purpose. The artifact will in this study be a model that helps companies combine internal and external data sources to improve their processes with the help of prediction models. The design and development will be done with the help of another framework from B. Bygstad and B. Munkvold in their paper: In search of mechanisms. This phase will be used to answer sub-question two, three and four.

When the model is completed it will be demonstrated in the next phase. This demonstration will be done for both companies(Lumen BS and Van Eijck). Because the model is tested at these companies, the demonstration will be a case study. During this demonstration it is important that the produced model works effectively and solves the problems that the companies under study have. This phase will partly answer sub-question five.

Based on this demonstration an evaluation will be held. With the help of experts at Lumen BS and Van Eijck, it will be evaluated how well the model performs. During the demonstration is will be observed how the model supports the desired solution, experts can afterward give their opinion on the artifact. With this feedback, we can iterate back to step 3 and improve the design of the model. This phase answers sub-question five.

The last phase is the communication phase. In order to give this research more meaning a chapter will be written on the importance and contributions of the proposed theory and model. With the help of example situations and the previous case study, the effect of this study on further works will be shown.

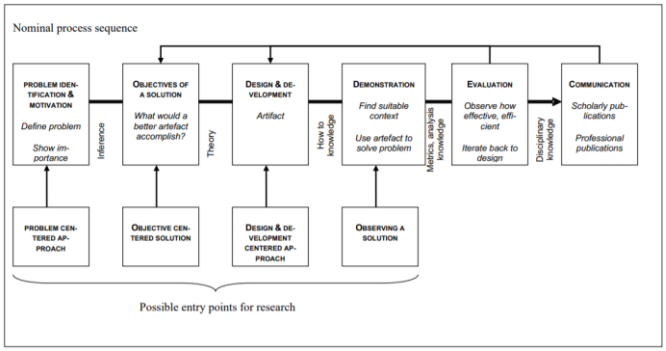


Figure 1: Design Science Research Methodology(DSRM) (K. Peffers et al, 2007)

For the literature research in chapter 2 a method by Kitchenham (2007) will be used. This is a well-established research method that exists out of 3 phases. These phases are: Planning, Conducting the review and Reporting the review(Dissemination). At the start of the literature review in the planning phase, the need for research questions will be discussed and a protocol will be developed in order to acquire the papers needed to answer the research questions. In the next phase, the actual research will be conducted following the set protocol, this means finding the papers that fulfill the requirements of the protocol and will thus benefit the literature review. The data from these papers will also be extracted in this phase in order to be used in the last phase. The last phase(Dissemination phase) exists out of reporting the findings that were extracted from the papers to answer the research questions. Once these research questions have been answered, there should

also be an evaluation to reflect on the conclusions of the paper and what its limitations are. By following these three phases of Kitchenham(2007) a well-structured and substantiated answer can be given on the research questions.

Like mentioned earlier, the DSRM will be combined with another framework in the design phase. The methodology that will be used during the design phase is a methodology produced by B. Bygstad and B. Munkvold in their paper: In search of mechanisms (Bygstad & Munkvold, 2011). Conducting a critical realist data analysis. In their research they suggest a methodology for information systems and have tested it on cases, their cases did also include data from CRM system. The main points from their methodology are that it is an improvement upon current empirical studies in the information system field by providing ontological depth, creative thinking and more precise explanations. The structure of the methodology looks as follows:

1. Description of events

2. Identification of key components

3. Theoretical re-description (abduction)

4. Retroduction: Identification of candidate mechanisms

5. Analysis of selected mechanisms and outcomes

6. Validation of explanatory power

Using this methodology, important findings can be made on which improvements can be suggested to improve processes within the company under investigation and design the model. To give a clearer picture of the methodology that will be used, all steps will be elaborated.

Step 1: Description of events

The first step is to describe events that have occurred in the company under investigation. In a critical realist context events are clusters of observations, which may have been made by the researcher or by the researcher's informants (Sayer 1992). This information can be gathered by conducting interviews with the company to discover certain events. Examples of events are: the reason why a CRM system was implemented in the company, recent mergers of the company and traineeships that were held.

Step 2: Identification of key components

During this step the most important components of the CRM data will be identified. These components can be for example persons, organizations or systems. By identifying these components, causal relationship can be explained more easily. These components can come forth from the data in different ways. This can be in a grounded way(Volkoff et all., 2007) or they can be embedded in a theoretical framework(Danermark et al., 2002). An example of components that were found by using this methodology on data from a CRM system are: the company, the CRM vendor, the exchange relationship and a government knowledge transfer program.

Step 3: Theoretical re-description (abduction)

To be able to work with retroduction we need to abstract the case, exploring different theoretical perspectives and explanations (Danermark et al. 2002). This means in this case that an elaborated literature research will be conducted to gain knowledge on the multiple important factors that play a

role during the investigation of the data. By gaining this information, causal relationships can be explained more easily and in a more meaningful way.

Step 4: Retroduction: Identification of candidate mechanisms

According to the researchers behind this methodology, this is the most important step. Retroduction is the opposite of deduction. With deduction, facts are formed based on hypotheses and with retroduction, hypotheses are based on facts, in this case from the data of the CRM system. Because this step is the most important one, it has been split into two sub steps.

Sub-step 4.1 The interplay of objects: Using the objects which have been identified in step 2, mechanisms can be found between these objects which are usually socio-technical. Mechanisms are in this case causal relationships between different objects. To investigate the interplay of the objects, it will have to be investigated how different objects interact with each other and if this interacting causes the desired outcomes for the company. An example which is given is a research from Lyytinen and Newman (2008) which used the four elements from Leavitt's diamond (people, technology, organization and tasks) to describe how the interplay between them constituted the mechanisms of socio-technical change.

Sub-step 4.2 Looking for micro-macro mechanisms: According to DeLanda(2006) there are two different types of mechanisms that have to be investigated:

- The micro-macro mechanisms: which explain the emergent behaviour, i.e. how different components interact in order to produce an outcome at an overall level for the company.

- The macro-micro mechanisms: which explain how the whole enables and constrains smaller parts in the system. In this case, the intended use of micro and macro is that object are micro if the relationship with higher entities in the companies is being investigated and they are macro if a relationship with an smaller entity is being investigated. Using this techniques a complete picture of an objects its causal relationships is being created.

Step 5: Analysis of mechanisms and outcomes

The next step in the methodology is analyzing the mechanisms that were found in the previous step. During this analysis, the focus should be what the triggers of the mechanisms are. A way to do this analysis is by performing the Context-Mechanism-Outcome form (Pawson and Tilley 1997). The outcomes from this analysis can then again be analysed using forward chaining,this way the intentions of the mechanism can be found or use backwards chaining, to understand the results of the mechanisms.(Pettigrew 1985).

Step 6: Validation of explanatory power

The final step is to look back at the found mechanisms and find the key mechanisms that influence the vital processes in the company, because the aim of the methodology is not to find as many mechanisms as possible but the most important ones. The key mechanisms can be found by investigating which causal structure explains best the observed events. To determine this the information gathered from the literature research must be used. When the key mechanisms are found that hinder processes, a tool can be found to adjust the mechanism in such a way that the company processes are being optimized.

Limitations of this methodology are that for it to work enough knowledge of the field must be available. Enough theoretical insight and domain knowledge must be available to perform step 1 and 3. During this research, the knowledge will be available due to a literature research and information

provided by Lumen BS and Van Eijck group which have a lot of domain knowledge. Using the information provided by them and using this approach a rich and precise set of explanations can be found on the CRM system's data set. After which these explanations can be used to optimize processes within the company especially when it comes to the time it takes for employees of the salvage company to arrive at the incident

1.6 Structure of the report

In this section the structure of the study will be described. This will be done on chapter basis. All the research questions will also be mapped to the chapters in which they are answered. Table 1 shows which chapter answers which research question and the methodology that is used.

The structure of the research in the report looks the following:

- Chapter 1 includes the problem identification, motivation, the research context, research questions and methodologies used during the research.
- Chapter 2 presents and discusses existing knowledge on system analysis. The knowledge that is gained during this literature research will be used to create an approach for using external data in combination with internal data.
- Chapter 3 describes the stakeholders related to this project with the help of interviews. By gaining this information of the stakeholders, more context is generated for the research and using the information gathered from the stakeholders the approach for External data intergration and analysis will be produced.
- Chapter 4 shows the design method for creating value for Van Eijck and Lumen BS by finding data connections between the data from the CRM system of Van Eijck and the external weather data that is added to the database of Van Eijck in this chapter. The results of this case will be used for the creation of the approach for using external data in combination with CRM-data.
- Chapter 5 describes the approach for using external data in combination with external data based on the literature, interviews and case study. This approach will exist out of multiple steps which will all be described.
- Chapter 6 evaluates the approach which has been created in chapter 5. This will be done using questionnaires and a group interview with experts in the field of system-analysis.
- Chapter 7 presents the discussion of the results from this research and the recommendations to Van Eijck and Lumen BS on how they can build on the results of this research. The limitations and future work are also described in this chapter.

Research Question	Research Methodology	Report
RQ1. What is the state of art of process optimization using system-generated data?	Systematic Literature review	Chapter 2
RQ2. What are the elements which should be focussed on when improving process optimization using system-generated data?	Systematic literature review	Chapter 2
RQ3. What is the current best-practice in industry to find connections between internal and external data?	Semi-structured interviews	Chapter 3

RQ4. What goals do the stakeholders of the Van Eijck case of optimizing processes using weather data have, which is used for the development and evaluation of the external data integration approach?	Semi-structured interview Literature review	Chapter 3
RQ5. How can we integrate external data and analyse it based on the case study in such a way that it takes less time and effort than is currently needed for the integration and analysis?	Critical realist data analysis	Chapter 4
RQ6. How to design an approach that fits the goals of the main research question based on current literature, experts and the case study?	Critical realist data analysis, Literature research, Interviews	Chapter 5
RQ7. How can the usability of the proposed approach be improved based on an evaluation by experts?	UTAUT Semi-structured interviews	Chapter 6
RQ8. How can the proposed approach help organizations in general improve their processes with external data compared to existing external data integration approaches?	Literature review, semi-structured interviews	Chapter 7

Table 1: Mapping of research questions to chapters

Chapter 2. Literature review

The literature research has been conduction following the guidelines of Kitchenham (2007). Existing out of 3 phases. These phases are: Planning, Conducting the review and Reporting the review(Dissemination). The purpose of this systematic literature topic is to create a clear overview of existing approaches on system-generated data. By creating this overview and looking at the strong and weak points of these approaches, future data analysts can easier choose which approach they want to take when analyzing system generated data and the overview of the approaches can be used to generate new models based on the benefits of existing ones. To create this overview of the data approaches the central research questions on this systematic literature review are: **What is the state of art of process optimization using system-generated data?** and **What are the elements which should be focussed on when improving process optimization using system-generated data?**

2.1 Research process

To perform the conducting phase of Kitchenham(2007) multiple databases were used to acquire relevant papers. The used databases are:

- ACM Digital Library (http://portal.acm.org).
- Science Direct – Elsevier (http://www.elsevier.com).
- Taylor and Francis (http://www.tandfonline.com).
- Scopus (https://www.scopus.com).

In order to find papers that will contribute to answering the three research questions the following search terms were used: ((“CRM-system” OR CRM W/1 system) AND (“model” OR “process optimization” OR “analysis”)). By using this combination of terms models for data analysis and process optimization were found for the literature research. Another search term was: (“CRM-system OR CRM W/1 system) AND (“Data mining”) And (“Model”)). This query focussed more on the search data analysis part of the paper. These queries have been used in all the databases mentioned earlier in the paper to find answer the main research questions by answering the sub-questions first. The first search using the queries discussed earlier resulted in 41 papers which seemed relevant to the literature review. By looking at the papers, it was concluded that the most relevant papers regarding CRM-system approaches came from 2002-2020.

Database	Number of Papers left after criteria
ACM Digital Library	21
Science Direct	10
Taylor and Francis	6
Scopus	4

Table 2: Number of papers from each Database

To include only the most suitable information into the literature research a set of inclusion and exclusion criteria were formed to assess the current papers. The criteria were the following:

Inclusion Criteria:

I1. The paper discusses its found CRM approach based on a research of a CRM-approach of which the CRM-system is a big part.

I2. The paper presents a model or an approach that is linked to CRM-systems.

I3. The paper presents information regarding process optimization in such a way that it is applicable to CRM-data approaches.

Exclusion Criteria:

E1. The paper is published before 2002.

E2. The found frameworks and approaches in the papers have not been tested in a correct way, which causes the proposed theory to be invalid.

E3. The paper does not go into depth about the different aspects of its approach or model.

E4. The paper is a duplicate of a paper that was found in earlier research on one of the other databases which were used for the research.

E4. The paper was not written in English.

E5. The paper is not peer-reviewed

I1 is important for the inclusion of papers because some papers seemed at first to focus on CRM-systems but later turned out that their view on CRM within companies did not include a CRM-system. I2 means that not only the right CRM focus should be addressed in the paper but that the paper must also present a approach or model from its theory to be usable for the ranking of the CRM-models towards process optimization. I3 is proposed to ensure that theory behind process optimization should not necessarily be directly linked to CRM-system data but is formulated in such a way that it is possible to apply the theory of the paper on CRM-systems.

Regarding the exclusion criteria, E2 made sure that papers that did not present a viable testing method were not included in the literature review, because without a proper testing method the viability of the theory in the paper can only be assumed. E3 was added to the inclusion list to make sure that sub-question could be answered. If a paper did not go into depth enough about its framework or approach is would be difficult to determine its strength and weaknesses based on other literature. Some papers had also an English title and abstract which made them seem to be a good addition to the papers for the literature review, but during a further inspection the actual paper was not in English making it difficult to use the information within the paper. This is why E4 was constructed.

After reviewing the papers on the inclusion and exclusion criteria a smaller set of 15 papers was left from the four databases. Most of the usable papers were found in the ACM Digital Library. Whenever there was a duplicate found, it was counted towards the first database in which it was found. Table 3 shows the exact number of papers from each database.

Database	Number of Papers left after criteria
ACM Digital Library	7
Science Direct	4
Taylor and Francis	2
Scopus	2

Table 3: Papers, Database ratio

In order to have a clear overview of the project, these papers were divided into three subjects, which each represented one of the sub-research questions. The following subjects were chosen: CRM-data models, Process optimization and Success factors. Table 3. shows which paper was assigned to which subject. Some papers covered multiple subjects and are thus represented multiple times in the table. The models that were used for the research originated from: China, India, Iran, the Netherlands and the USA.

Subject	Reference
RQ1. CRM-data models	Reinartz et al. (2004); Li et al. (2019); Song, Haihong, Zhao and Zhonghong et al(2016); Valmohammadi et al(2017); Rygielski, Wang and Yen et al(2002); Gupta and Verman et al(2009); Bahari and Elayidom(2015); Engel & Schoonderwoerd(2020).
RQ2. Process optimization	Akroush et al.(2011); Battor & Battor, (2010); Keramati et al. (2010); Eldon Y. Li & Russell K.H. Ching (2009); Samaaranayake et al(2009); Markerink et al(2016).
RQ3. Success Factors	Rygielski, Wang and Yen(2002); Bahari and Elayidom(2015); Hugh Wilson , Elizabeth Daniel & Malcolm McDonald (2002); Alshawi, Missi and Irani(2010); Kambatla, Kollias, Vipin and Grama(2014).

Table 4: Mapping of references

2.2 Results

To answer RQ1. *What is the state of art of process optimization using system-generated data?*, the approaches and models that remained after the inclusion and exclusion criteria in the research method should be discussed on their way of approaching CRM-data. In the following section, the eight research models will be summarized on their shared qualities which form the current state of art of process optimization using system-generated data. After this summary a more detailed explanation of each model can be found.

2.2.1 Summery of researched models

The researched models for process optimization using system-generated data show that in the field of CRM-data models there is a difference in what the models focus on. Three of the eight models focus solely on the technical aspects of CRM-system generated data optimization, two models focus only on the business side and the other three papers look both at the business and the technical side.

The models that focus on the technical side cover the same three aspects for getting the most out of process optimization, these are: adaptability, data cleaning and repeatability. Adaptability is part of every technical model, data cleaning and repeatability are discussed in almost all technical papers. Adaptability means, how many different kinds of data formats a system can handle. The more data formats the system can understand the higher the level of adaptability is. High adaptability is important for process optimization using system-generated data because data from inside an organization can come from a lot of different sources with different data formats, all these formats should be understood by the process optimization system to get a full picture of the current state of the organization and how it can be improved.

The data cleaning aspect of the papers in which it was found looks at the level at which the system is able to transform and delete data that is inserted wrong. The data that is inserted in the system can have human errors and to improve the quality of the analysis the system should have the option to

find errors in the data and transform them to the correct standard or else delete the wrong data. The better the data is cleaned, to more reliable the findings from the analysis are which will increase the chance for a successful process optimization using the results of the analysis.

The repeatability which is covered in almost all technical model in process optimization shows the level of degree that the analysis of the system is understandable and can be repeated by other people and get the same results. If the system is to complicated and the results are not acquired by means which are understandable to the user of the system then the results are less reliable and the process optimization is more likely to fail.

The five papers that included the business side of process optimization using system-generated data discussed mostly the following two aspects: KPIs (key performance indicators) and business analysis. Especially the business analysis, which looks at the current state of the business before the actual process optimization analysis was important since it was covered in all five papers with business models. The inclusion of KPIs was discussed in two of the business papers, most organization have already established their KPIs, the companies that have not done this should according to the models produce these KPIs because these will be used to determine what to measure when performing the current state analysis.

The five aspects discussed above are the most important aspects of process optimization using CRM system-generated data in current literature and represent the state of art. A full explanation of each model and the importance of the aspects found in them is given in the next two sections 2.2.2. and 2.2.3.

2.2.2 Models for CRM data analysis:

Reinartz, W., Hoyer, W. & Krafft, M. (2004)

The first found CRM analysis model is the earliest constructed model found in the literature research by Reinartz et al. (2004). This paper presents a model for the performance outcomes of CRM-system process which existed out of three main parts: CRM Process, Economic performance and moderators. Reinartz argues that to measure the success of the CRM-system, you should not only look at the perceptual economic performances like most research did at the time, but also assess the association with a measure for objective economic performance (Varadarajan and Jayachandran 1999).

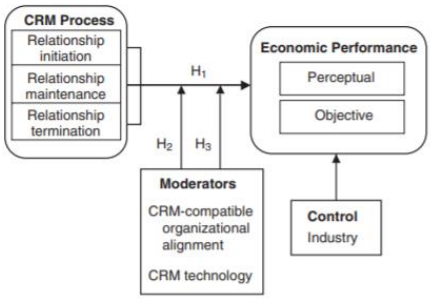


Figure 1: Model by Reinartz et al.(2004)

The CRM Process consists of three dimensions, which are: relationship initiation, relationship maintenance and relationship termination. These dimensions give a clearer depiction of how the CRM-system is used by different companies. To determine the effect of the CRM-process on the economic performance, there are also two moderators. The first one is CRM-compatible organizational alignment, which included training procedures in the company and employee

incentives to use the CRM-system. The second is CRM technology, which is the level of investment that has gone into the CRM technology for the company.

To test the validity of the model it has been tested using data from 1015 companies from Austria, Germany and Switzerland. The data was acquired using a survey. From this survey, it was concluded that the moderators have a significant effect on the usefulness of a CRM-system in a company. It was also found that the long term relationships and relationship initiations have a significant positive effect on the economic performance of the company

Li, Y., Huang, J., & Song, T. (2019)

The next model is proposed by Li et al. (2019). The paper proposes a model for the value of data in CRM systems by looking at IT/IS usage theory and "two-stage model".

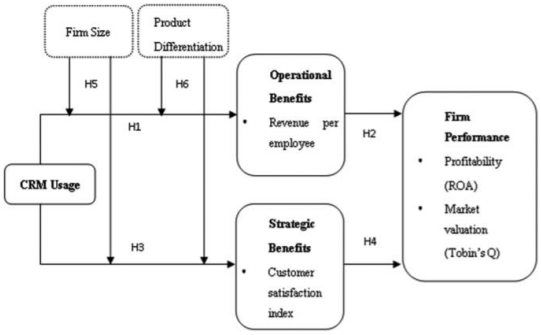


Figure 2: Model by Li et al.(2019)

Using the IT/IS usage theory and "two-stage model" they suggest that CRM usage combined with firm size and product differentiation are important factors to take into account when analyzing data of a company that makes use of CRM-system, because these factors influence the operational and strategic benefits that are generated by the system.

This model was tested by making use of Harte-Hanks CI Technology Database, Compustat, and ACSI as data sources. From this database were 378 samples gathered which has as requirements that they had to be from the fortune 1000 companies in the United States and make use of CRM. By looking at the introduction of CRM-systems in the companies and aspects of the company like stock value and customer satisfaction conclusion were made.

The empirical research has shown that in order to get the value of the operational benefits from the CRM-data, you should look at the revenue per employee and the strategic benefits are reflected in customer satisfaction. The correlation between the size of the firm and operational and strategic benefits is positive when CRM -systems are being used.

Song, Haihong, Zhao and Zhonghong (2016)

Song et al(2016) argue that the RFM model(Recency-Frequency-Monetary model) is a good starting model to analyse CRM-data. This is because the data from CRM-systems can be divided into internal and external data which have relationships both among users and characteristics. The RFM model is ideal to find information in the data about recency, frequency and monetary. The recency is the freshness of a certain activity in the data. The frequency is the number of times the activity happens and the monetary is total or average money that is made with the activity. Using this model, the most important activities can be classified and activities that underperform can be looked at. However, this model works best on small data groups and has not yet been proven to work on large

data sets. That is why Song, Haihong, Zhao and Zhonghong et al(2016) propose a new approach for data analysis of CRM-data, a multiple statistic-based CRM approach via time series segmenting time interval of RFM. This approach makes use of time series to divide the data into smaller parts which are easier to analyze with the RFM model. After this step, MCA model is used to analyze the relationship of several categorical dependent variables. The MCA analysis is done on two aspects: inner relationships of three dimensions in RFM model based on inner numerical characteristics, interaction relationships of these numerical characteristics and qualitative ones.

The model of Song et al(2016) has been tested using a dataset from a telecom service. This test showed that the model of Song et al(2016) is a viable methodology to analyze large data sets such as the data from a CRM-system. Both the internal and external aspects of the companies could be improved using the result of the multiple statistic-based CRM approach via time series segmenting time interval of RFM which Song et al included in their model.

Valmohammadi (2017)

The next research model for CRM-system-data is presented by Valmohammadi (2017). The framework that is presented in the paper investigates the relationship of the data from the CRM-system and organizational performance and innovation capability. CRM practices are divided in five subgroups which included the CRM-system. The framework is tested in a case-study on 211 Iranian manufacturing companies with the use of structural equation modelling.

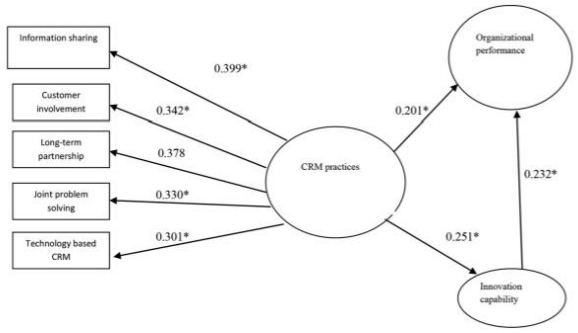


Figure 3: Model by Valmohammadi(2017)

From this case study it was concluded using this model it could be determined that the use of CRM-systems had a small positive correlation with the organizational performance and innovation capability. This was mainly due to innovation which was introduced with the CRM usage. This theory is not new and already suggested by Lin et al.(2010) who linked innovation with the use of CRM. This study did however confirm the theory. From this paper it can be determined that during analysis of the data generated by CRM-system the focus should be the relationships of innovative aspect in the company with the organizational performance. The organization's performance can be measured by return on assets, return on investment and profit margin on sales, sales growth, market share, market share growth, customer satisfaction and overall profitability (Akroush et al., 2011; Battor & Battor, 2010; Keramati et al., 2010). To see if changes in organizational processes are beneficial, this data should be analyzed.

Rygielski, Wang and Yen(2002)

Rygielski, et al(2002) provide an approach with two alternative options based on what kind of analysis the user wants to analyze data from a CRM-system, the neural networks model and CHAID(CHI-square Automatic Interaction Detector).

The first option is the neural network model is a model originally from NeoVist Solutions, Inc. This solution of gathering useful information from CRM-system data makes use of pattern discovery tools based on neural networks, clustering, genetic algorithms association rules.

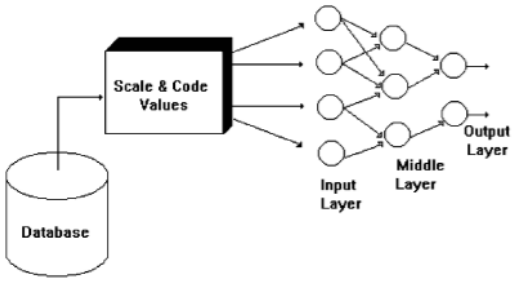


Figure 4: Model by Rygielski et al(2002)

The second option is CHAIN this is used to give companies a competitive advantage by optimizing sales and marketing productivity through segmentation modeling. The focus of the model is to maximize the lifetime of customers and acquire new customers at a low cost. Through the use of CHAIN a predictive model can be produced, which makes it easier to make future decisions based on the data from the CRM-system.

Overall the neural networks model can be wider used than CHAID because of being able to be applied to both supervised and unsupervised data mining. Neural networks can also handle categorical and continuous independent variables like the status of a project and total income better. However, CHAIN is more useable for exploratory problems instead of estimation problems, due to being able to provide descriptive rules. CHAIN is also easier to use on a data set which makes it more user-friendly and a cheaper option than neural networking. The ease of use comes from the fact that the neural network approach works more like a “black box”, it comes with predictive solutions but it is hard to explain how the outcome is determined. This is not the case with the CHAIN model, which like mentioned before has great explanatory power.

Bgattacharya, Godbole, Gupta and Verman(2009)

Bgattacharya et al(2009) argue that the optimal approach for analyzing data from a CRM-system is by using an approach that is asset-based, repeatable and adaptable. During their research, they developed an approach called IVOCA. IVOCA stands for e IBM Voice Of Customer Analytics and is a hosted asset-based, managed service offering for CRM analytics. IVOCA makes use of five phases to perform its analysis. The first phase is the gathering of data using data sources. Next is data processing & conversion stage, here data is separated into structured and unstructured data. Following is the data storage stage using Indexed files and IBM DB@ warehouses. Once the data is stored, it is analyzed in the analysis stage and after this, it is reported in the reporting stage.

During the making of IVOCA, they discovered that the aspects: asset-based, repeatability and adaptability were highly important for the success of the tool. The tool that is used must be able to seamlessly utilize data from various data sources. Being able to get useful information out of unstructured data is of importance to be able to use as many data sources as possible.

The repeatability was another important aspect for the approach towards a CRM-data-analysis tool. In this case, the repeatability is not only that the analysts are able to repeat similar tasks on different data set but is it also important that if different researchers analyze a data-set, that they have the same outcomes. The outcomes of the data analysis should be reported in such a way that different researchers come to the same conclusion. This is done by making the predictive modeling flows as easy to understand as possible.

The final important finding is that the service provided should be able to analyze the data without being interrupted by the researcher during its complex analysis.

Understanding of the data should be high enough for the service to produce predictive models without the researcher filling in blanks in the system, which would also lead to different results between different researchers.

Bahari and Elayidom(2015)

Bahari aet al(2015) have created a CRM-data mining framework, which works efficiently to generate predictive models. This new framework does split the data mining into multiple phases. The first phase is understanding the business goals and requirements of the problem domain. This is followed by the data preparation phase, which includes data transformation, attribute selection and cleaning of the data. Using this prepared set of data a model is built in the model building phase, which is used to generate a prediction. This model is evaluated in the next phase and the last phase visualizes the data generated by the model.

The most important step of this framework is the model building phase. The paper argues that model building should exist out of: Classification, Association, Regression, Forecasting and Clustering.

The Model was tested using results of direct bank marketing campaigns from 17 different Portuguese banks. The data originated from 2008 till 2010 and contained 46211 instances. 10% of this data was used for the evaluation, which showed that the framework gave correctly classified instances enough times to be successful.

Engel and Schoonderwoerd(2020)

The final approach for CRM-data analysis is proposed by Engel et al(2020). They have used Garner analytics model to analyze how well data-analysis was used in 16 different companies that had at least a turnover of 500 million euros. Garners analytics model differentiates between four levels of maturity of the analysis. The four levels are: Descriptive, diagnostic, predictive and prescriptive analysis. The more complicated the analysis becomes, the higher the value of the analysis will be. A descriptive analysis does only say something about what is happening in the company. The diagnostic analysis also explains why a certain phenomenon is happening. The more complex variant of this is the predictive analysis, which will predict what will happen instead of saying what is currently happening. The highest valued analysis s the prescriptive analysis. This analysis tells the user based on given information how they can archive a certain goal within the company. The higher the information process optimization is, the higher the level of the data analysis can be.

Based on the interviews and the use of the Garner analytics model a set of best practices has been made for the use of data analysis. The first practice that is proposed is not technical advice but one on a managerial level. In order to get the most out of a data analysis within a company, a vision must be created by the managing board. The board should decide together with an external and internal expert what they want to achieve with the implementation of a data analytic tool. This decision should be about the Garner level and it should include what kind of data should be extracted from

the data analysis. To ensure that the success of the data analysis the first implementation should not be to be big. It is better to start with a pilot and build upon this than to start with a system that is to complex.

Key performance indicators should be made to evaluate the performance of the analytic tool. It is highly important that these KPIs are in line with the business vision and strategy of the company. KPIs should be formulated in such a way that they do not just something about one compartment of the company but the performance of the whole company. This will results in fewer KPIs, which is beneficial because it makes it easier to evaluate the KPIs. An example of a KPI that tells something about the whole organization is the change in total profit since the implementation of the analytic tool.

During the research of Engel and Schoonderwoerd it was found that almost none of the organizations under investigation had thoroughly thought about which tool they wanted to implement, but simply implemented something that worked at other companies. For the success of the analytic tool, an organization should not implement simply implement something that worked by others but look at their own business vision and KPIs that were made and find the tools that are best in line with these criteria. The paper advices to make use of Gartner Magic Quadrant for Analytics and Business intelligence Platforms to get a clear overview of whether a tool is right for the intended vision and implementation of the organization. However one of the main aspects which is important for almost all implementations is the ability of a tool to analyze data from different data sources.

The IT landscape of the organization must be flexible to realize the implementation of the analytic tool. The chosen analytic tool based on the vision of the organization should be able to be integrated into their current IT landscape. Big changes to their IT landscape should be avoided. Current providers of CRM-system have made their applications in such a way that they can be adjusted to the IT landscape of different companies, which gives organizations still a wide variety of choices when choosing an analytic tool based on their vision.

The last practice that is advised by Engel and Schoonderwoerd(2020) is to involve the employees during the decision and implementation of the analytic tool. The implementation of the analytic tool means that employees will work more fact-based and they should be prepared for this. A pilot is advised to see the reaction of the employees to the new system and let them get used to it. The research speaks of ‘coaltion-of-the-willing’ this means that if management is able to convince a group of employees that the new system is a good addition, they will become early adopters and other colleges will follow once they see the early adopters using the new fact-based analytic tool.

2.2.3 Process optimization using CRM-data analysis

The previous section has elaborated on important models regarding data analysis that are applicable to data from CRM-system. The focus of this research is on the usefulness of these models on data to improve business processes in organizations. To evaluate the models on their improvements on business processes it should first be established what the important aspects are for process optimization, to create the criteria for ranking the models of section 5A.

Qin et al (2014) formulates a set of important factors to effectively get improvements in process optimization using large data sets like the ones that are generated by CRM-systems, which will be used as criteria for ranking the models from section 5A.

According to the research data analytics is an indispensable tool for improving the key processes within an organization. Data can provide realistic information about unknown phenomena that are

occurring in the organization and can predict future phenomena based on current information. With this information, adjustments can be made in the organization to improve key processes. For example, the data might suggest that a certain department works less efficiently due to having not enough group meetings, then this can be adjusted in order to increase efficiency and thus optimize processes. The data analysis can also be used to monitor the effect of the changes that are made in the organization. If the changes are in line with KPIs that are made before the change, then the positive effects can be measured. The processes that can benefit the most according to Qin et al(2014) are high-level optimizations such as planning and scheduling because the manpower of a certain task compared to the results of the department is easily compared.

For the model or approach to be effective for process optimization it should be easy to understand and use. The more complex a system is the harder it is the implement and understand the data that is produced will also be harder to understand and thus be less usable for process optimization. Although employees should also familiarize themselves with the new system the threshold should still be low. The most important point of knowledge of the system is the data-processing architecture, knowledge about algorithms and the user should be able to judge the results of the analysis tool and not blindly trust the data. These are important aspects that a model must possess and suggest to be usable for data optimization.

Ungermann, Kuhnle, Stricker and Lanza (2019) have also found a set of factors that a model should have to be useful for process optimization. To be effective for process optimization an approach should first look at the current state of the organization before performing the analysis. This analysis should be well documented to be able to be used in the end. This is done through a current state analysis at the start of the research. The current state analysis should include the boundaries of the system, these include the current state of value streams, data landscape, process control loops and sensor technologies. To get the clearest depiction of the current state of the system it is best to divide the systems technologies into four layers, based on aggregation level. Models should have the possibility to include this analysis to get a depiction of what the focus should be during the data analysis and based on this create KPIs.

The model should have the possibility to include KPIs which are formed from the analysis of the current status of the organization. In order to optimize processes, the focus of these KPIs should be optimization. These KPIs are during the analysis linked with their mathematical relation if the model that is used works correctly, using a learning method. To create the right connections between the KPIs and the data, training data should be used at first to learn the model what kind of factor in the data should be connected with what indicator. This should be validated manually in the beginning to assess the correctness of the system. So while full automation is the goal, the analysis should include ways to also manually assess the connections of the system in a way that is not too complicated for the user to understand.

When looking at the data analysis it should include a clear data cleaning and transformation. For process optimization, the approach must take into account that data will come from multiple data sources to analyze different aspects of the organization which can lead to a better optimization. A correlation analysis is also of high importance to determine which factors should change to optimize a process.

To get the most out of the analysis, the results should be categorized into four categories: personal, organizational, material and equipment related measures. An optimal model would be able to place the found connection on its own in the right category, but in most cases this is done manually by people who have a lot of general information about the organization that is optimized. Once the

connections are placed in the right category another assessment should be performed in which the cost of improving a connection is made and also look at the potential benefits of making changes to the connection. Using this, for each connection a cost-benefit ratio is produced, which can be used to choose the optimal recommendation for the organization.

Gröger, Niedermann and Mitschang(2012) focused during their research on the important factor which data mining should include to be beneficial for process optimization using this data. They found that the two most important factors for the assessment of the models were the interpretability and the robustness of the approaches was.

Interpretability means the degree to with users can understand the decisions that have been made by the system. The higher this degree of interpretability, the easier it is for the user to understand the prediction that has been made by the system.

The robustness of the system, means how effective the algorithm of the system is in handling datasets. The more effective the system is at generating prediction models using different types of data the higher the robustness of the approach is.

For a model or approach to be effective for process optimization using data analysis the degree of interpretability and robustness should be high to get the best results. This means that the approach should be able to handle different types of data and still produce the correct connection between aspects of the organization and that the produced results are easy to understand for the user to determine where the certain connection came from and check their correctness.

Based on the aspects given in this section it can be determined whether the models from the previous section are useable models for process optimization.

The first model of Reinartz et al. (2004) is the oldest model discussed in the previous section has formed the basis for other models on how to work in a CRM way using CRM systems. While being a good starting point for the research, its main focus is CRM techniques and uses CRM technologies as moderator for success. Because the CRM system is just a moderator, the model is not useful for creating process optimization using CRM systems.

The model proposed by Li et al. (2019) could be useful for optimization because it gives a useful correlation between the usage of CRM systems and operational benefits which are produced by process optimization. The moderator, firm size is not mentioned in previous literature about data analysis and optimization but seems an interesting aspect to take into account when assessing the potential benefits of using data mining. While product differentiation can be led back to the Ungerman et al(2019) claim that the system should be able to determine outcomes using different kinds of data sets, this can be said the other way around as Li does. By keeping the product simple enough that different kinds of data sets are not necessary, the data mining becomes easier and better predictive models can be made for optimization.

Song et al(2016) is a useful model for process optimization because it makes use of internal and external data which is important for optimization according to Ungerman et al(2019). While most models focus on the effectiveness of internal data, it is key to also make use of external data to get a better depiction of the meaning behind correlations. However, the robustness of this model is low due to being only effective on small data sets with the time series. The time series do increase the degree of interpretability, but to be the most effective for process optimization both robustness and interpretability must be high according to Gröger et al,(2012).

The model proposed by Valmohammadi(2017) does like the model from Reinartz show the effect of CRM systems on process optimization but only uses it as a small part of the model. While it can be useful to know other CRM methodologies techniques for increasing the organizational performance, the nature of this research focuses on the data analysis of CRM systems and thus this model is not optimal to generating more optimizations only to show how much a data system could potentially benefit the organization while being combined with other techniques.

The two models provided by Rygielski et al(2002). These were less focused on the business side of process optimization and more on the approach of the data mining which can be used to create the predictive models. The two approaches were CHAID and neural networks combined with CRM system data. Both models show potential according to the factors mentioned by Gröger et al,(2012). Neural networks have a high degree of robustness, but like mentioned before are like a 'black box' when it comes to the way it makes decisions and thus has a low degree of interpretability. CHAID is the other way around, it has a high degree of interpretability, but its robustness is lower than neural networks. Of the two approaches, CHAID should be the preferred one for process optimization, because the interpretability is of high importance for good results and its robustness is still okay, although lower than neural networks.

Bgattacharya et al(2009) made use of their IVOCA model which is useful for process optimization because of multiple factors. The three main aspects that they discovered and made the focus of the IVOCA model are: being asset-based, repeatable and adaptable. The repeatability is an important factor for process optimization according to Qin et al(2014). This is in the sense that multiple people within must be able to make use of the analysis model and get the same results.

The adaptability of the IVOCA model makes is also useful for data cleaning and data transformation which are important aspects of process optimization according to Ungermann et al(2019).

This model does however have a downside and that is that it makes use of a complex analysis method to produce the predictive models. This causes the interpretability to be low for the IVOCA model. This means that while it is easy to perform the research understanding what is happing in the model to produce the predictive model is hard. This reduces the usability for process optimization analysis Gröger et al(2012).

Bahari et al(2015) framework is the first model investigated that also includes a pre-phase to determine the business goals of the organization. This is extremely important for achieving a successful process optimization Ungerman et al(2019). Their model also makes use of a preparation phase to transform and clean the data, which is also one of the important factors discussed earlier to achieve process optimization. In the next phases, the model is evaluated and visualized. This means that if the model can be evaluated that the analysis technique has a low degree of interpretability. Due to being easy to understand keeping in mind the goals of the organization and being able to process different kinds of information, the Bahari et al framework is a great framework for process optimization using CRM system data.

The last evaluated approach is the one by Engel et al(2020).

This approach focuses like the framework of Bahari et al(2015) on the inclusion of KPIs when performing an analysis and not just focus on random connections which are purposed by the predictive model. Another positive aspect of the framework is that it mentions that the tool that the company will use should be tailored to the organization and more importantly being able to analyze data from different data sources. The ability to use data from different data sources increases the degree of robustness making the approach a better fit for optimization.

The model from Engel et al(2020) is the only one that mentions the alignment of the chosen approach and the organization's landscape. It focuses on getting the employees involved during the process which is important according to Ungermann et al(2019) to be able to get the most out of an analysis. With the help of employees, it is easier to categories the results in the four subcategories that Unermann et al(2019) purposes for process optimization, personal, organizational, material and equipment related measures. This combination of points makes the approach of Engel and Schoonderwoerd a great fit for process optimization using CRM data according to current theories on optimization.

2.2.4 Ranking of models based on process optimization

Based on the evaluation in section 5B a ranking has been made of the usability of the models for process optimization using CRM data. Although not all models focused on the same aspects, some focused more on the technological side and others on the organization side, the usability could still be measured based on current theories on process optimization, explained in the previous part. The criteria that were used for the ranking were the aspects of process optimization suggestion by the papers in the previous section. These models were ranked based on their business level of optimization, their level of technical contribution to optimization and the integration of the business and technical side. The criteria for having a good business level for optimization are the inclusion of KPIs and business analysis like a current analysis, which will show the status of an organization before the CRM-based changes are implemented. The level of technical contribution to optimization was assessed by the inclusion of important optimization techniques like: adaptability, data cleaning and repeatability. Based on these criteria and the analysis from the section 5B the models scored as following and are ranked by that score:

Ranking of models based on Process optimization	Model	KPIs inclusion	Business analysis	Adaptability	Data cleaning	Repeatabilty
1.	8). Engel et al(2020)	+	+	+	+	+
2.	7). Bahari et al(2015)	+	+	-	+	+
3.	5). Rygielski et al(2002)	-	-	+	+	+
4.	6). Bgattacharya et al(2009)	-	-	+	+/-	+
5.	2). Li et al(2019)	-	+	+	-	-
6.	3). Song et al(2016)	-	-	+	-	+
7.	4). Valmohammadi(2017)	-	+	-	-	-
8.	1). Reinartz et al.(2004)	-	+	-	-	-

Table 5: Ranking of models and performance

2.3 Strong and weak points of models:

Based on the finding from the previous section a set of points have been determined which are currently done well on the model on CRM-analysis and a set of weak points have been found which should be tackled by future models to get the most out of large CRM-system data sets.

2.3.1 Strong points of current models:

The current models showed promising results for process optimization both on the technological and the business side. One of the aspects that is currently in almost all the inspected models is the importance of adaptability. The ability of an analysis tool to be able to not only inspect one type of file but multiple and also be adjustable to learn other file types is of great importance according to the literature and almost all current models of data analysis for process optimization do stress also stress this. Only the papers that focused solely on the business side of data analysis did not include this.

Another point that is of great importance for process optimization is the repeatability. This means that if someone performs the analysis multiple times, the same result will be presented and more important, if different people perform the analysis it also keeps the same results. To realize this the degree of interpretability should be high. Almost all the models that were investigated provided a model that made it easy for the user to understand what was happing with the data once it was inserted. This means that the degree of interpretability is high and thus process analysis can be easier achieved.

Data cleaning is according to multiple papers key for getting the most out of data sets like CRM-system data. Most papers did show this in their model. This is a strong point for the models because without the initial data cleaning, the predictive models become less reliable or the system could get errors.

The most important point for the technical and the business side of the models is only included in the first two models for the ranking but is a very strong point for both these models to get the most out of process optimization using CRM-system data. This is the inclusion of KPIs as starting point of the model. The inclusion of this in the top two models made them stand out from the rest because according to the literature KPIs are key to achieve the goals of the organization. KPIs create a clear overview of what the organization wants to accomplish improvements and make it easier to decide which correlation of the prediction model to pursue.

Besides the KPIs, the inclusion of a current state analysis is a strong point of the two top models. These current state analyses can be made used to generate the KPIs, because by carefully analyzing the organization, the researcher will find what the key processes are in the company and how they can be measured. Besides being useful for the KPIs they also provide useful documentation to use once changes are made in the organization, to see if the changes truly improved the organization compared to the current state.

2.3.2 Weak points of current models:

By reviewing the models for CRM-system data analysis using current literature on process optimization it became also clear that there is room for improvement for the models to become more efficient.

Section part 1 of section 5D showed that a lot of models have a high degree of interpretability, but to get the most efficient analysis the level of robustness should also be high. This combination of a high/high degree in interpretability and robustness is nowhere found in current models. Gröger et al (2012) looked at other approaches outside this research and found the following results.

Classification Technique	Interpretability	Robustness
Decision Tree Induction	High	Low
Bayesian Classification	Low	High
Decision Rules Generation	High	Low
Neural Networks	Low	High
Support Vector Machines	Low	High

Table 6: Classifaction Techniques assessment

The model shows that not only the approaches investigated in this research but classification techniques in general only have a high degree in one of the two aspects. To improve this a technique should be found that is not only easy to understand but also is highly effective in handling datasets.

Although the involvement of KPIs to optimize processes are present in the top three models from the model ranking (Table 5: Ranking of models), most models did not include them. This is a weakness in most models. The inclusion of the KPIs would be extremely beneficial for the organization that is using the model. This weak point relates directly to another that the models investigated for the research focussed primarily only on either the business side of CRM-data analysis or the technical side. To produce the most useful model for companies the model should have a hybrid of these two. An approach that shows how a business can not only decide what the goal of the analysis with the help of KPIs are for the organization but also present the best focus points for the technological part of the analysis would be highly beneficial. Currently, organizations have to find their own blend between these two subjects, which results in either the business or technological side not being represented enough in the final solution.

When we do focus on the business side, one of the weaknesses of the models is that the employees are not getting involved enough in the optimization process. Employees have a lot of knowledge about specific parts of the organization and should therefor be included as an essential part of starting phase of the analysis when generating KPIs. By including the employees in the starting phase of the optimization, the transformation phase of the organization becomes easier. Because due to their involvement, employees are more willing to accept changes that will be made in the workplace for optimization.

No current model in this study does have a clear categorization of its results once the predictive model is produced. To increase the strength of the models, the correlation in the predictive model should be categorized this makes it easier to implement the changes. Besides that is also is useful to use in a final analysis to see with correlation category had the most positive impact and which categories could still be improved.

2.4 Conclusion

This research focused on two questions: RQ1. What is the state of art of process optimization using system-generated data? and RQ2.What are the elements which should be focussed on when improving process optimization using system-generated data*?* RQ1 was answered by first looking at approaches found in the literature. These approaches suggested business approaches to get the most out of system generated data and there was a set of papers that primarily focused on the technological side of CRM-system-generated data. These papers were evaluated using a set of points from literature based on process optimization to review how well the approaches were usable for optimization within an organization using a big data set like CRM-system data. From this evaluation a ranking came forth (Table 5: Ranking of models) which showed that the model by Engel et al(2020) was best fitted for optimizing an organization using system-data, due to its combination of business

and technological factors. The focus of this model was also on the alignment of the approach with the company which was a big strong point.

RQ2 was answered using the found strong points in section 2.3.1. These strong points were found using the ranking (Table 5: ranking of models). The most noticeable strong points of current approaches are the integration of a data cleaning phase before the actual analysis, making sure that the data approach was understandable for the user and the adaptability of the system to process different file types. These factors were the technological strong points of most papers. Especially the repeatability was a point that came forth from existing literature and the models. This shows the importance to know what the system is doing and not blindly follow the outcome of the analysis. The business side was mostly not addressed in papers but the ones that did gave a good inclusion of the importance of KPIs and a current state analysis.

To increase the strength of the models the weak points of current models should be overcome to improve the current papers for CRM-system data-based optimization. These weak points were that although most models had a high degree of interpretability, this is in no case combined with a high degree of robustness. Models should search or researchers should try to develop a classification technique that has a high/high degree when it comes to interpretability and robustness, meaning that the technique is powerful but it is also understandable for the user how the system developed its prediction model. Another weak point that integrates with the interpretability is categorizing the results of the analysis to make it easier to understand and implement is hardly mentioned in the researched papers

Besides these most important technical improvements, there are also business improvements for the current models. The comparison of process optimization literature and the models showed that models do not integrate the role of employees enough during the process of optimization using CRM-date. Including this in models will increase the chance of finding correlation to optimize and makes it easier to make changes in the workplace. The inclusion is at the moment still too limited. Only two of the investigated models made use of this while it is of high importance for successful process optimization.

This leads to the main weakness of current models, that the blend between the business and the technological side is minimal. Current models have shown that they only focus on one of the two aspects while a blend between the two would be optimal for organizations to realize process optimization using CRM-system data. In this blended model, the best pre- and post- phase should be discussed to find KPIs and how the changes can best get implemented and the blended model should show what the optimal factors are for the actual analysis to get the best prediction model. Using the combination, the success chance of optimization within an organization would increase.

2.4.1 Limitations

One of the limitations of this research is the amount of investigated approaches. A reasonable amount of CRM related models have been found, but there might be more models which were not included due to the timeframe. Other models could in the future also be added to the ranking of CRM-models based on the literature on process optimization. This is also the case for the papers on process optimization.

Another limitation is that the models included in this research were only models that were published in English papers. There could exist usable CRM models in other languages that would improve this research but due to the language barrier they could not but included. If researchers would like to try

and expand the knowledge on CRM systems they should publish in English, for this is the language of scientific conferences.

While all the papers were in English, they were performed in different countries. A lot of countries were not included in the empirical researches of the investigated papers. The models investigated for this research originated from five different countries mentioned in the research method section. Due to this underrepresentation of countries, the results might not apply to all regions in the world.

Chapter 3 Stakeholders

This chapter will focus on the most important stakeholders for the research. It will explain what their motivation is for a successful completion of this research at what their stakes are. This will be done with the help of interviews with experts in the field of data analysis and a stakeholder analysis.

3.1 Stakeholders

Stakeholders have been identified by their contribution to the research and the effect that is will have on them. The identified stakeholders are: Van Eijck, Lumen BS, Rijkswaterstaat and Companies that provide weather data(MSN, KNMI). About each stakeholder an explanation will be given what their stakes are for this research.

Van Eijck

Van Eijck is a salvage company in the south of the Netherlands. This means that whenever a car breaks down in their region they are sent to the incident to help the person in need. This can be done by repairing the car or if the car is not fixable, by towing the car away. Van Eijck has multiple control rooms which send out cars to help people in need and assess the situation. At some incidents a safety car has to be provided by Van Eijck and in other cases this is done by Rijkswaterstaat. Van Eijck does not only help passenger cars but also bigger vehicles like trucks. Whenever vehicles have to be towed away they also provide a place where the car can temporarily be stored.

In may 2020 Van Eijck decided to implement Microsoft dynamics 365 into their organization as CRM system. A CRM system compiles data across multiple channels on how the organization interacts with clients. These are the different points at which the company interacted with the client and how they have interacted with the client. The goal of this system is to improve how the organization approaches and works together with the client to drive sales growth. This can be done by automating standard tasks using the CRM system which have been identified. For example automatically filling in forms, based on earlier information. The CRM system also provides information on how much revenue is generated from each client. This gives the user of the CRM system the option to compare revenues between different quarters easier. The CRM system makes it also easier to generate reports on how the company is performing since most of the information about the company can be found in the CRM system. The difficulty of CRM systems is to train your employees in using the system correctly. To get the most out of the CRM system the employees have to insert data into the system how they have handled the client. The format of the data should be taught correctly to the employees to avoid errors in the data.

The reason that Van Eijck has chosen to implement a CRM system in May 2020 is that their current systems were outdated. No new updates were provided for their system which if nothing would have been done could lead to security risks and a competitive disadvantage because the system is not as good as it could be. Van Eijck is also growing so the managers decided it was time for a new system that could provide them with clarity on how the organization is performing. To see how the company was performing they decide to implement Microsoft dynamics 365 with the help of Lumen BS. They chose this system because it allowed Van Eijck a lot of possibilities in configuring the system to their needs. The system also constantly updates which keeps the system safe, new and stable. With the help of Lumen BS the CRM has been completely configured to the needs of Van Eijck and they now have control over their own data and can easily see how the organization is performing. In most CRM systems customer retention is one of the goals of the system. Obviously, Van Eijck does not want to retain the customer that has had an incident, but another type of customer is the government which assigns districts to Van Eijck in which they are allowed to work. They want to

keep these districts, otherwise they lose work space. The official term for the districts in which a salvage company is allowed to work is a rayon. Besides retaining these rayons, the CRM system makes it easier to do tenders, based on the financial information provided by the system. The schedules are also made using the CRM system, because based on historical data it can be determined how busy it will probably be on certain days in the week.

Soon after the implementation of the CRM system, Van Eijck extended the usability by adding Power BI to their company. Power BI makes it easier to view the data from the CRM system and perform analysis with the data from the CRM system. Power BI is an extension on the CRM system, used to view and get more info out of the data that is gathered. During this research analysis will be performed in de BI environment of Van Eijck on their CRM data.

For this research Van Eijck was chosen as the test case to learn more about the implementation and analysis of external data in a CRM system. Van Eijck wants to learn more about the outliers in their data. These can at the moment not always be easily explained. The external data source that was chosen is weather data. This is the first time that Lumen BS has helped with inserting and analyzing external data at one of their clients. That is why it was chosen to pick a dataset that has a clear connection with the current data of Van Eijck and was obtainable for free. Based on the experiences from this research Van Eijck can choose to insert more complicated data into their database and make use of bigger but paid options. This could be information that Rijkswaterstaat has on the quality of the roads and where the incident has happened if the road has multiple lanes.

Lumen BS

Lumen BS is a partner of Microsoft that is located in Zoetermeer and Nijmegen. They provide their clients with advice about CRM and ERP systems and help them implement these systems. Their goal is to let their clients work more efficiently using their advice and systems. Together with the client they will sit together and see what their needs are and how a Microsoft Dynamics 356 ERP or CRM systems could be tailored to the needs of the client. One of these clients is Van Eijck, they have been provided with a CRM system tailored to their needs by Lumen BS.

Lumen BS also has as goal to constantly improve and innovate itself. That is why they decided to work along with this research. They want to be able to provide their customers with more options to work more efficiently. One of these options is helping clients make use of external data set to explain their data and find relationships that can be improved upon. This is why this research will create an approach on how to implement, analyze and transform an organization with external data. This will be done through the use of literature, knowledge of experts and the experiences from the Van Eijck case. This approach will be used to help their client and can be built upon. The goal of Lumen BS is to create a standard method on how to help clients best based on previous cases, by knowing what to focus on and how to analyze is the easiest.

Rijkswaterstaat

Rijkswaterstaat is the executive part of the ministry of infrastructure and water authority. They are part of the government and are responsible for the main roads and waterways. Since Rijkswaterstaat is responsible for the roads they give out the contracts to the different salvage companies to work in certain rayons. Whenever an incident has happened the salvage company is contacted directly by the person who had an incident or the salvage company is sent to the incident by Rijkswaterstaat. The goal of Rijkswaterstaat is to make the traffic go as smooth as possible by keeping the roads in a good condition and resolving traffic jams whenever these happen.

Rijkswaterstaat wants the salvage companies to arrive as soon as possible at incident locations and resolve them. This is why they benefit from this research if salvage companies are able to arrive sooner at incident locations because of improvements that have been made in the organization based on the use of weather data. If Rijkswaterstaat notices the positive effects of external data on the salvage companies they are more likely to work more closely together with salvage companies when it comes to data sharing. Data that Rijkswaterstaat could provide to salvage companies could for instance be the quality of the roads and a description on which side of the road the incident has happened. These could in the future help salvage companies to arrive even sooner at incidents.

External weather data providers(KNMI, MSN Weather)

The external data for this research will be as explained earlier weather data. This data is a good starting point to develop further from. The companies that provide the external data are KNMI and MSN Weather. They provide web services that make it possible to add weather data to your database if they can be connected with a time, longitude and latitude. The goal of the weather providers is to provide their customers with the most accurate weather information. Information from their website is free but there are also paid options. This means that though showing the quality of the data in the free version they want the client to use the paid version.

3.2 Stakeholder analysis

Based on the information that has been gathered from speaking to the stakeholders and information online, the stakeholders from section 3.1 will be mapped with the help of a stakeholder analysis. This will provide more context to the research by showing what the goals and influences of each stakeholder are. In this analysis each stakeholder will be rated on the impact and influence on them at the project and how they should be treated.

Stakeholder	Impact	Influence	What is important to the stakeholder	How could the stakeholder contribute to the project	How could the stakeholder block the project	Strategy for engaging the stakeholder
Van Eijck	High	High	Van Eijck wants to find out what the influence of the weather is on the incidents on the road and see if it can explain the outliners in the data.	By providing information on the data of the CRM system which makes it easier to create the visualizations needed for the external data analysis.	If they deny access to their CRM system database the project is no longer possible.	Keep Van Eijck up to date on what the results are of the analysis with regular meetings to see if the analysis is still going how to want it and to see if no wrong assumptions have been made about the data.
Lumen BS	Medium	High	They want to learn from	Lumen BS provides	Lumen BS could deny	Regular meetings

			the results of the research, what the best practice is when it comes to performing analysis with external data to use the knowledge at other clients.	information on how to start and perform regular analysis with clients based on their years of experience in the field.	giving help making it more difficult to perform the business and technical analysis without their expertise.	should also be planned with Lumen BS to keep them updated on the progress and ask questions if needed.
Rijkswaterstaat	Medium	Low	Rijkswaterstaat wants the roads to be safe and without traffic jams. When a person has an incident, help should arrive as soon as possible.	In the future Rijkswaterstaat could provide external information to enrich the database of Van Eijck even more. They also provide knowledge on the salvage sector which is needed for the project	If Rijkswaterstaat stops giving Van Eijck the contracts for their rayons, then, no data comes to the CRM system anymore and the project cannot continue.	By letting Van Eijck be able to arrive on time at incidents, Rijkswaterstaat will stay pleased and will not disrupt the project.
Weather data providers	Low	Medium	Want to provide clients with the best possible data on the weather and sell this.	Provide Weather data that can be connected with the data of Van Eijck to perform the analysis.	If the provider decided to stop providing the data, the external analysis could continue using another weather data provider, however although this is possible it is not desired, since data can be saved with different values, this	Weather providers do not have to be kept engaged. Their web services can be used without their engagement.

					will take time to convert to the right standard.	

Table 7: Stakeholder analysis

This stakeholder analysis shows that Van Eijck and Lumen BS are the most important stakeholders and the highest combination of influence and impact. The influence and impact of the stakeholders can further be analyzed using the power-interest matrix for stakeholders. For this matrix the influence corresponds with the power and the impact with the interest of the stakeholder. The stakeholders are mapped on the following places in the power-interest matrix.

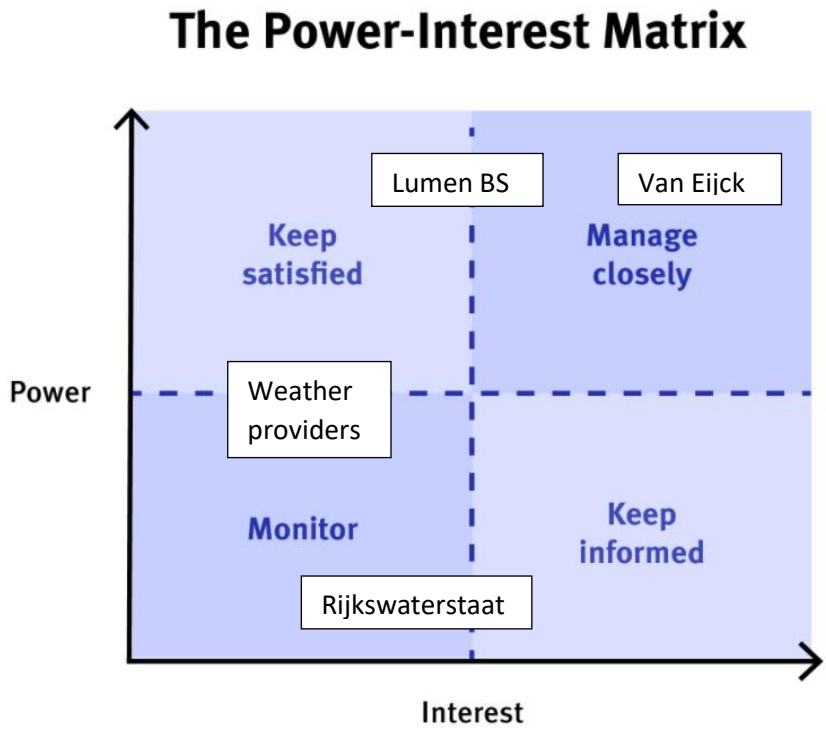


Figure 5 Power/Interest matrix

As Figure 5 shows, Van Eijck must be managed the closes of the stakeholder because they have the most power and interest of the stakeholders. This will be done by performing regular meetings to see if the results still meet their expectation and if there have not been any mistakes when it comes to data elements. After this Lumen BS should also be managed closely. Their knowledge of the system and business side of the project is key to get a good result for Van Eijck. This result will also make it possible to make a better approach towards external data integration at the end of the project. Rijkswaterstaat and the Weather providers are important stakeholders for this project but they are not directly influenced by the results of the project. This is why it is not necessary to manage these stakeholders closely.

3.3 Interview Data analysis

To answer the second sub question related to connecting internal and external data and get a better overview of the interests of stakeholders and get a better idea of how some stakeholders look at the

use of external data, interviews have been held with two experts in the field. These answers will help during the actual analysis in the design phase. The questions focussed on the connection between internal data and external data and the properties that the data must have in order to be usable for this analysis.

This interview was a semi-structured interview with a technical business consultant and an expert on data analysis from Rijksoverheid. It has been prepared using Galletta, A: Mastering the semi-structured interview and bey: From research design to analysis and publication (2013). The interview questions and the answers can be found in the appendix.

Based on the interview an analysis has been performed to find the most important factors for analysing internal data with external data. The results from this interview have been coded using the technique from Saldana, J. (2012). The Coding Manual for Qualitative Researchers. SAGE Publications. The goal of coding is to symbolically assign summative, essence-capturing attributes to the language based data that has been acquired with the interview. This is done by first processing interesting statements of the interviewee into an initial coding, which shows the essence of the statement. The next cycle is to group the initial coding statements into focused coding groups. These focused coding groups are produced out of the overhauling themes of the initial coding. The focused coding shows the most important results from interview. In this case the focused coding will exist out of the focus points for when external data is combined with internal data for an analysis. The coded focused and initial results from the interview are the following:

Focused coding	Initial coding
Data Cleaning	General importance data cleaning
	Standardize fields for data to be clean
	Relationship quality of data and data cleaning
	Do not let the database be larger than needed
	Too much data will lead to messy data bases
	Importance of clean data towards analysis
	Reduce option to get cleaner data
	Inspect data on errors
Organizational changes	Instruct employees on use of system
	Show importance of data
	Train people to check data
	Created general understanding
Ease of use	Choose based on knowledge
	Do no restrict options
	System should fit into company
	Keep the system simple
Price	Simpler systems provide cheaper options
	Clients focus on price tag
	Managers implement changes based on profit expectancies
Contact client	Find out what client wants
	Find the focus points for the analysis
	Inspect analysis with client
	Keep the client connected during the analysis.

Table 8: Interview coding

From the coding the following main categories were found: Data cleaning, Organizational changes, price, ease of use and contact client. These are the five focus points when organizations want to analyse their company using a combination of internal and external data and will be further discussed in section 3.4.

3.4 Conclusion

The goal of the interviews was to answer RQ3, what is the current best practice in industry to find connections between internal and external data?. Based on the analysis of the interviews, the first thing that should be focussed on for the connection between internal and external data, is the quality of the internal data. The quality should have a high standard in order to be usable during the research, if the quality is not good enough, then a large part of the database will become unusable, which decreases the validity of the findings. If the validity decreases because the database gets smaller, the chance is bigger that causal relationships are not significant and purely accidental. To increase the quality of the data, the data must be structured and clean. It can be kept clean by standardizing the input fields. The way the data is saved is also important for the clearness of the database. All parties should agree on the saving format of the data, otherwise, the data will become vague. An example of this is that a location field that has Amsterdam written in it, could be about Amsterdam in the Netherlands, but also about Amsterdam in America. This should be avoided, an solution for the example problem would be using latitude and longitude, this format is accepted around the whole world. Another thing that is important about data saving is that only data should be stored which has a purpose for the database. Do not start saving everything just because it is possible. If everything is saved, then this will results in an unclear and slow database, with too much data. Wanting to save everything will also increase the workload on your employee, which will result in more errors. This is due to the more the employees have to fill in for the database, the higher the chance is that they make mistakes.

Standardizing and agreeing on what kind of data should be saved is not the only way to increase the quality of the data. There will probably remain fields that cannot be standardized and be filled in by employees. To increase the chance that the data that they fill in is correct, The employees should be informed on the importance of the data and what the data will be used for. Still, if they know about the importance, the manual inserted data should be checked on peculiarities. If these peculiarities are found they should be transformed or deleted to improve the database. If someone deletes or transforms data in the database it should be recorded, to avoid people falsely changing the data, to make the company look better than it actually performs.

Once the data is transformed and cleaned, it can be used for the analysis. To combine internal and external data, there are multiple ways. The analyst could use special databases, with time/space analysis and a rule engine that on the fly the inserted data checks. This is an older way of analyzing data but is still usable nowadays. It does not present the data in a fancy way, but this makes the solution really cheap. The analyst can also still choose to use an off-the-shelf solution in the form of Microsoft Power BI. This product is still relatively cheap and easy to learn. It can be used to analyze a combination of internal and external data as long as Power BI does have to make a request every time it needs its data. The external data should be saved in a table and be updated from time to time, to be usable in Power BI. This updating frequency should however not be too often but relative to what the data is used for and which frequency is needed. If the internal database is connected with the external database, Power BI can be used to visualize how the external factors influence the internal processes. To increase the power of Power BI, Azure can be used together with the Power BI system. This way prediction models can be generated, that show the organization what the future will roughly hold for them if they keep working how they do.

Due to the partnership with Van Eijck for this research, it has been chosen to use Power BI as the analytic tool. Power BI has already been implemented at Van Eijck by Lumen. This means that the internal databases are already created and filled to be used for the analytic part of this research.

Chapter 4 design method

The design method has been performed using the critical realism approach consisting out of: a description of the events, identification of key components, a theoretical re-description, a reproduction, analysis of the mechanisms and outcomes and a validation. Each section will contain one of the steps from the critical realism approach starting with the description of events in section 4.1. By following these steps the data analysis can be performed in a structured way, increasing the chance to find interesting correlations in the data.

4.1 Description of events

This section will describe the three events that have occurred at Van Eijck which are important for the data analysis, because they influence the data or are the reason the data exists. These events are the implementation of the CRM system and Power BI, a recent merger and the corona virus. Providing a more in depth explanation of these events will provide context for the data. These events were found with the help of literature online and interviews with Van Eijck and Lumen.

The first event is the actual implementation of the CRM system and a small explanation of the company which is being analysed. The company under investigation is Van Eijck. Van Eijck is active in the salvage sector in the south of the Netherlands. In May 2020 they have implemented a CRM system from Lumen BS. This is the Microsoft dynamics 365 system. This system has been implemented to create a database with information about the incidents that have occurred and how they were handled. In October 2020 Power BI was also implemented at Van Eijck, Power BI was implemented to create reports on how the company is performing, the goal of Van Eijck is to use it more and more as a way to find connections in the data that show where the organization should change to improve the data in the CRM system instead of only showing how the organization is currently performing.

The second event is the merger of Van Eijck. Van Eijck has establishments in Hulten and Eindhoven, together with dependances in Breda, Den Bosch, Nistelrode, Raamsdonksveer and Valkenswaard. Last year they started a cooperation with Van Amerongen. Van Amerongen has establishments in Arnhem and Ede. The goal of this cooperation is to grow faster in the salvage sector in the Netherlands and internationally perform more repatriation activities. Van Eijck will focus mostly on the international market in the cooperation.

The third event had not only big consequences for Van Eijck but also the rest of the world, the Corona Virus. February 27th the first person in the Netherlands was found to have the Coronavirus. In March the first lockdown was announced by the government. Due to the lockdown, people had to work from home as much as possible. This lead to fewer cars being on the roads, which lead to fewer cars being involved in an incident. Recent studies from CBS show that last year since the introduction of the lockdown, the number of incidents has reduced by 23 percent. The lockdown measurements were mostly removed in June and July, but due to the second corona wave, a second lockdown started at the end of 2020. The CRM system of van Eijck was introduced during the first lockdown, so a difference in incident numbers cannot be shown in this data. The data from the CBR does however tell that fewer incidents have happened, which means that Van Eijck will be busier again once government regulations are lifted and more people start going to work again. This means that the results from this research apply to the lockdown situation. If more people go to work again and it gets busier on the roads, the recommendation from this research might not apply.

The employees of Van Eijck work also from home if possible. The drivers of the incident cars can still do their normal jobs while taking the corona measurements into account. The people in the control

room could also still go to their jobs. This means that the way of working has not been massively disturbed by the coronavirus when it comes to sending the cars to the incidents. On the management side things have changes since meetings have become online meeting and the creation of rapports based on the information from the CRM system is now done at home and not at one of the buildings of Van Eijck.

4.2 Identification of key components

There are a couple of components that are integral for the CRM system of Van Eijck. By identifying these key components, the possible relationships between these components can be explained more easily which also increases the chance that a relationship is found which can be used for the process optimization. These components could be persons, organizations systems or part of the system. The key components have been identified in a grounded way (Volkoff, O., Strong, D. M., and Elmes, M. B. (2007)). This means that based on the data and knowledge that has been acquired by interviews and a tour through the current CRM-system the key components were identified of Van Eijck. Other key elements that will be explained are the external data and the systems which provide this data. The research of the thesis focuses on the use of external data which makes the external data systems clear key elements. Knowledge about the external data has been acquired with the help of Lumen BS and online literature.

4.2.1 System

The first one is the employees of Van Eijck who go to the incidents on the highway. These employees go to the incident to pick up the vehicle that has broken down and take it with them. It takes a certain amount of time for an employee to arrive at the incident. The time the incident is notified to the employee and the time it took to arrive at the location is saved in the CRM-system and is used to assess how well Van Eijck is performing. Employees can insert it into the system when they arrive at their location and also add other data, like the car type and number plate. The inserted data by the employees are added to the database in the form of an incident report. Each individual incident has its own report. This way it can easily be investigated if an employee has done something wrong while handling an incident and conclusions can be drawn on the performance of the organization by combining all the information of the incident reports. The CRM system also keeps track of the invoices that have to be sent to the clients and the invoices that Van Eijck still has to pay.

4.2.2 rayons

Another key component of Van Eijck are rayons. These rayons are certain sections of roads in the Netherlands. The Netherlands has 221 rayons as can be seen in figure 5. Each rayon has its own savage company that takes care of the incidents that happen on that part of the road. Van Eijck has roughly 40 rayons. Savage companies can compete against each other to get hold of a rayon. This is done by showing that you can remove the defective cars from the roads the fastest. During negotiation, savage companies tell the divider of the rayons how fast they can retrieve the cars on average and the savage company that has the lowest time and is big enough to handle the rayon, gets the rights to service in the rayon.

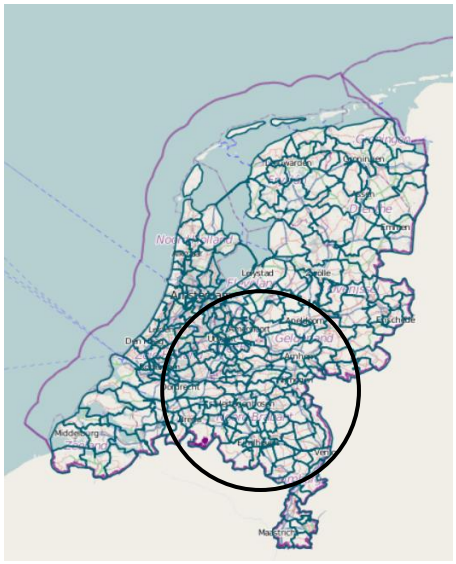


Figure 6: rayons in the Netherlands with highlight of Van Eijck rayons

As mentioned in the previous paragraph the rayons are divided based on how fast a savage company can retrieve a car from the highway. This average time is put into a contract and the savage company needs to perform according to this contract otherwise they will lose the rayon. There is a one-minute marge on the contract time, but if the average becomes too high, the rayon is lost and other savage companies can bid on it again. It might be the case that the same savage company gets the same contract again, simply because they have no competitors in the rayon. The time that the savage companies propose is not the only guideline that they have to follow. There are also rules that apply to all rayons. These are that during day time, which is from 05:00 – 23:00 the savage company should arrive within 20 minutes and at nighttime, which is from 23:00 – 05:00, the employee should arrive at the incident within 25 minutes. 90 percent of the trips to an incident should be in accordance with these times otherwise the rayon is also put up for bidding again. The measurements of these rayons are done on a quarterly basis. Van Eijck does not want to lose its rayons, so it is of utmost importance to know how they are performing at each rayon and if they find a rayon that does not perform well enough, fix the problem.

4.2.3 Services

The different kind of services that Van Eijck provides is also key. In total Van Eijck provided ten different types of services: *Vehicles that er left behind, assistance, tax authorities, direct transport, incident management, salvage, roadside assistance, coordinated transport, heavy salvage and tow away*. The first service, *vehicles left behind*, means that Van Eijck is sent to pick up a car that is left behind in an obstacle-free zone. Cars left in this zone can cause dangers for the other drivers, that is why they to be picked up.

The next, *assistance* means that vehicles of van Eijck will be on standby somewhere, where a call is expected or that they drive preventive to a location, where an incident is expected. If an incident does happen and they are on standby, then the time it takes to arrive at the incident is really low, which is beneficial to the average time that cars have to drive to an incident in a rayon.

Van Eijck also provides services for the *tax authorities*. When people can no longer pay or do not want to pay the tax authorities, their vehicle can be taken into custody. Van Eijck will pick up the car and store it until the bill to the tax authorities is paid off.

Direct and coordinated transport is like the transport of a vehicle from location A to B, which can have multiple reasons. This can be direct, in this case the transport is mostly done within an hour after the order has come in or it can be coordinated. With coordinated transport does not have to happen immediately, but can be done once Van Eijck has time free in the next few days.

Incident management is the measurements that have to be taken to keep the roads safe. This does not apply to all roads but to the so-called IM-road, which exists out of the highways, a part of the provincial roads and a part of the municipality roads.

Salvage or heavy salvage applies when Van Eijck does not only have to keep the road safe where the incident happened, but also has to take the car with them because it is in no condition to drive anymore. The car will be taken to one of the depots of Van Eijck for further steps.

Tow away, means that the car has to be taken to a different location by Van Eijck after the incident, but it is not stored at Van Eijck, but brought to another safe location.

The last option is *road assistance*. This is an important service for this research because Van Eijck is aiming to improve this one. Road assistance is provided if the car has broken down but is still fixable with the help of Van Eijck. If the car turns out to be in a worse state than expected, then the road assistance can be turned into tow away or salvage.

4.2.4 External Data

The goal of this research is not only to look at the internal data of Van Eijck in the form of the incident reports but also to combine it with an external data set, which will hopefully explain certain aspects of how Van Eijck is performing. The chosen data set for the external data is weather data. This set is chosen because it is expected based on experience that the weather has a high influence on the number of incidents that happen on the roads and it possibly influences the time it takes for employees of Van Eijck to arrive at the incident. Van Eijck also already had contacts which could provide them the weather data fast, which was an important requirement due to the scope of the research. In the future Van Eijck could start implementing more external data into their database. This could be data from Flitsmeister and Rijkswaterstaat. The info that could be extracted from their database could be on which side of the road the incident has happened. This is important information for the safety of the road and helps Van Eijck when they send out a safety car to an incident to be better prepared. The side of the road on which the incident has happened can also give explanation on why some incidents took longer to handle than other incidents. Another option could be getting more information about the quality of the road from Flitsmeister. By knowing which roads are in a poorer state than other, more incident cars can be deployed near them.

To get the weather data a connection has been build between the database of MSN weather and the database of Van Eijck. It has been chosen to use the MSN weather database, because it provided a lot of information about the weather, was free to use and relatively easy to connect with the Van Eijck database with the help of web services. Weather in the database of MSM is saved with time, longitude and latitude. The incident reports of Van Eijck also include time, longitude and latitude. This way every incident report is filled with weather data. This data is from the time that the employee of Van Eijck has arrived at the incident. This means that the weather might have been slightly different when the actual incident happened because it takes an employee roughly 15 minutes to arrive at the accident. The connection between the database of MSN and Van Eijck has been made at the start of April 2021. This means that since April the incident report also has weather data added to them. MSN weather does not provide weather information from earlier instances. This data includes the following: temperature, wind chill, dew point, windspeeds, gust

speed, wind direction, air pressure, humidity, UV-index, UV-index description, Weather condition, air condition and visible distance. The weather condition is based on a set of factors from the rest of the information to determine whether it is for example cloudy, sunny, partly sunny, rainy etc.

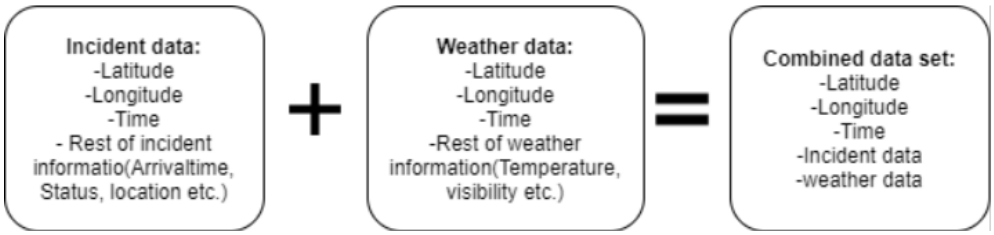


Figure 7: Combining incident data with weather data

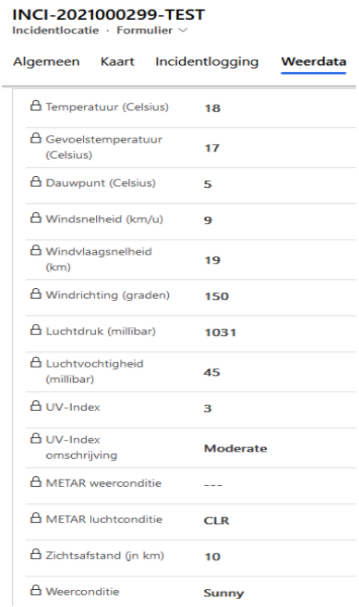


Figure 8: MSN weather data connection

To make the data set larger it was chosen to include another database that has less weather information but does safe historical data. This is the database of the KNMI. A downside of this database is that the historical data is saved on a daily basis and not hourly as the new incident reports have. This means that all the incidents that happened on a day before April 2021 will have the same weather information. The other downside of the KNMI data is like mentioned earlier that the data set is less complete than the data from MSN weather when using the free option. The paid option has not been used because the client first wants an impression of the additional value of the smaller external dataset, before implementing more external data by paying for it. The historical data from the KNMI does not include the wind chill, dew point, UV-index, UV-index description, weather condition and air condition. Especially the lack of weather conditions is a bit of a shame, that it was decided to determine the weather conditions ourself based on the rest of the information that was provided. This way the older data was still usable enough for the research. An overview of the database used for the research including the weather data is given in figure 10, here the weather data is already inserted into the incident table.

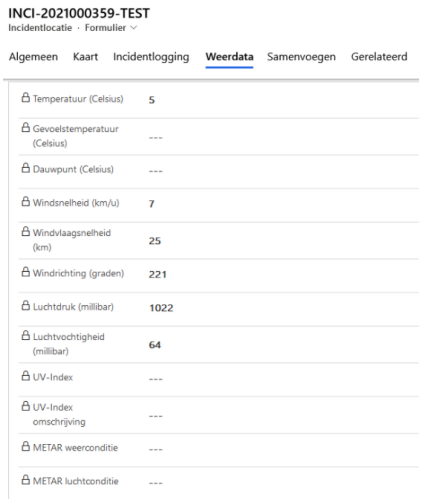


Figure 9: KNMI weather data connection.

The exact descriptions of the weather data points are:

Name	Path	Type	Description
`Pressure`	responses.weather.current.baro	float	The atmospheric pressure.
`Conditions`	responses.weather.current.cap	string	A caption of weather conditions such as rainy, sunny, etc.
`Dewpoint`	responses.weather.current.dewPt	float	The temperature at which dew forms.
`Apparent Temperature`	responses.weather.current.feels	float	The apparent temperature, or feels-like temperature.
`Humidity`	responses.weather.current.rh	float	The relative humidity percentage.
`METAR weather conditions`	responses.weather.current.wx	string	The METAR code of weather conditions.
`METAR Sky Conditions`	responses.weather.current.sky	string	The METAR code of sky conditions.
`Temperature`	responses.weather.current.temp	float	The current temperature.
`UV Index`	responses.weather.current.uv	float	The numerical UV index.
`UV Index Description`	responses.weather.current.uvDesc	string	A description of the meaning of the UV index.
`Visibility Distance`	responses.weather.current.vis	float	The visibility distance.
`Wind Direction`	responses.weather.current.windDir	integer	The wind direction in degrees clockwise from north.
`Wind Speed`	responses.weather.current.windSpd	float	The wind speed.
`Wind Gust Speed`	responses.weather.current.windGust	float	The wind gust speed.

Figure 10: description weather data

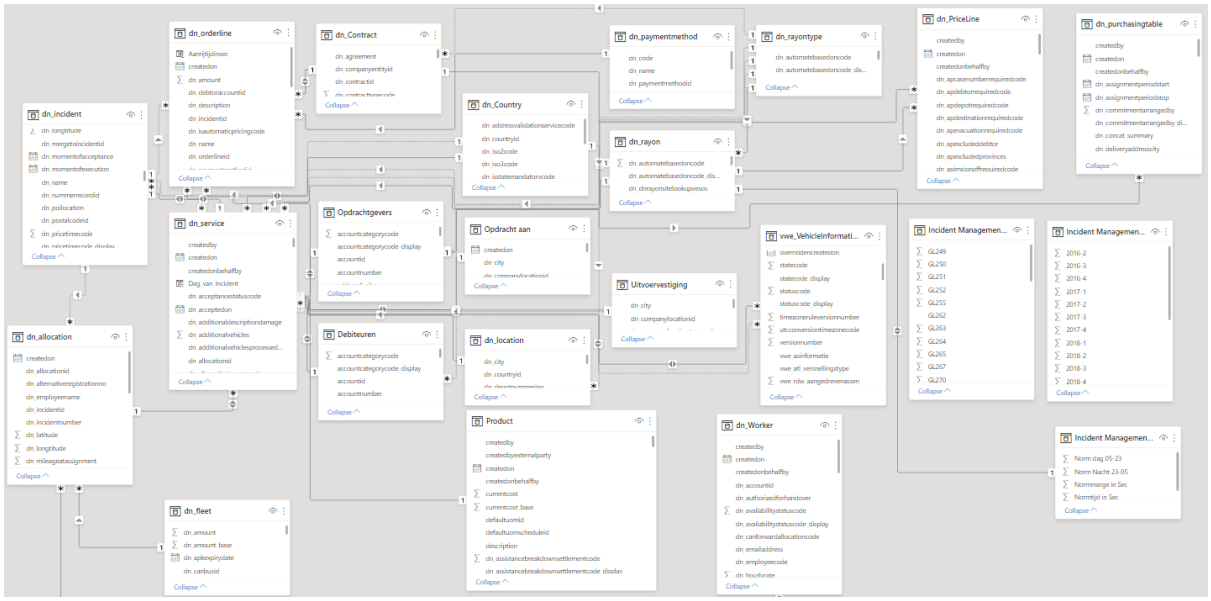


Figure 11: Relationship tables in database

4.3Theoretical re-description (abduction)

To analyze the data a deeper understanding of external data use must first be created with the help of a literature research. The critical realism method states that we need to abstract the case, exploring different theoretical perspectives and explanations (Danermark et al. 2002). However due to there being no usable literature on the salvage sector and the focus of this research being on external data, it was chosen to dive deeper into the literature behind the use of external data. This is acceptable when looking at the critical realism method because the theoretical re-description should be on a case that transcends the actual events of the research organization. In other words, the theory from this subsection should not only be applicable to Van Eijck but also other companies. This section will provide more information based on literature on how to acquire and use external data. This helps providing more context to the case of Van Eijck and is also usable at other companies when they want to implement external data.

The first paper that was looked at is: Other People's Data, Petschulat, S. (2010). This paper shows how organizations can effectively integrate external data into their own databases. According to the paper it is important for every organization to make some critical decisions based on external data. There are multiple ways external data can be acquired. Flat file data feeds, web services and webpages can be used to fill the data warehouses of an organization with external data. A lot of web services and data mashups are available online which have data about the weather to retail sales data. Organization should take full advantage of these available data sets.

External data can have different kind of structures. They can also not all be accessed in the same way. The most detailed level of external data is called the base data. This kind of data is well structured and is the easiest to understand. The most common way that this base data is acquired is using flat files which are transferred with a file transfer protocol.

Another Method for acquiring the data is making use of webservices. Webservices are quite popular and information from for example: currency, current events, the weather and competitive intelligence can be called using these webservices. The added value of using webservices instead of flat file data transfer is that it is easier to combine the webservice with other services. These could

be calculations, filtering or a search function on the data set. This makes the data clearer and improves the analyzing process. The functions allow the data to be cleansed, which is an important data aspect for data that will be used in an analysis according to the results of the interviews in chapter 3. There is however a downside to the use of webservices. They are only useful if the creator of the webservice has implemented a query for the information that you need. For example the EBay webservice does provide a query for a lot of calls, like top vendors, products or highest bids. But if you need something specific from this database that Ebay has not thought of then the web service is not the solution and a flat data file might be more usable.

There is another method that can be used and that is using data found directly on the web itself. This data is unstructured and can be accessed using web pages and search engines. Information from the web itself is in most cases combined with the data sets of webservices. This combined information set makes it easier to explain findings during the data analysis. The method of simply looking the information up from the internet and putting it in excel files is called web scraping. This method has to be done manually and is therefore more expensive than webservices, it is also error-prone because it is done manually. There might however be cases that the creation of a new webservice takes too much time and data structure changes too often. In this case they might choose to use web scraping to get a competitive advantage.

When choosing to make use of external data three things should first be assessed. These are the flexibility, quality and cost of the external data. Flexibility refers to how easy the data is to use for the purposes of the organization. Before acquiring external data sources an organization should investigate what the what kind of data they need. If this is found out than a choice can be made how they will acquire it. Raw flat files have the highest likelihood of being usable for the purpose of the organization because it is highly adaptable and can be transformed in many different ways. This does however mean that a lot of time will have to be put into transforming the raw file, to fit the needs of the organization. Web services are a lot more summarized, but like mentioned earlier might not have the correct calls in them to fulfill the needs of the company. The amount of flexibility the company wants the data to have determines the method that should be used.

If the type of data has been determined and a method has been found with the right flexibility, the quality of the data has to be assessed. The organization that provides the data has to be looked at and the type of flows that will be used to transfer the external data to your own database. Some of the data might be extracted using formulas and it is wise to check these. The paper, Petschulat, S. (2010) has compiled a set of questions to look at to see if the quality of the data is good enough to extract. These are:

• Do all of the formulas use the exact same algorithm?

• How do they deal with rounding errors?

• Have currency conversions been applied?

• Has the data been seasonally adjusted?

• Are nulls treated as zeros or a lack of data?

The last aspect but certainly not the least important aspect for most companies is the costs. Not every organization has the same budget. Bigger companies can in most cases afford to spend more than the smaller ones. This does however not mean that smaller companies cannot make use of external data set, but they have to make trade offs on the quality and flexibility in order to get the data at an affordable price. The quality of data can also be maintained in house if you choose to

make use of a cheaper option with low quality data. This does however mean that man-hours have to be spent on improving the quality. This price should be lower than the money that has been saved by choosing a dataset with a lower quality. You can save on flexibility if you create queries in house that could also be directly bought. Here the same applies, costs of labor should not exceed the savings of the cheaper option. Having everything in the core data warehouse at the lowest possible level of detail while still being usable represents the extreme of maximizing flexibility and quality while trading off cost. While looking at the costs it should also be kept in mind that the more data is saved the higher the maintenance costs will be, so do not safe to much information and too frequent. These points were also brought forwards during the interviews from chapter 3. For the costs is also a set of questions produced by the paper of Stephan Petschulat, to determine which trade offs can be made to reduce the costs for the external data, while keeping the value of the data for the company.

• What is the business impact of incorrect data?

• What is the cost of maintaining the data feed?

• How large are the datasets?

• How often does the data change?

• How often does the data schema change?

• How complex is the data?

• How complex and varied are the consumption scenarios?

• What is the quality of the data (how many errors expected, how often, magnitude of impact)?

• How critical is the data to decision making?

• What are the auditing and traceability requirements?

• Are there any regulatory concerns?

• Are there any privacy or confidentiality concerns?

If the right external data set and method have been found based on flexibility, quality and costs it is time to look at the next step. This is when the external data should be integrated into the architecture. The architecture of the organization has multiple layers and inserting the external data into one of these layers has its own benefits and drawbacks. The different stages are during ETL(extract-transform-load), in the core data warehouse, at the data marts, during consumption with BI-tools and in the application.

The earliest stage to enrich the data is on the way in. Using a vendor the incoming data stream can be cleansed, transformed and augmented when it is being sent to the data warehouse. Because of these transformations, the data that gets in is of high quality. The incoming data set is often quite large, this is why for this earlier stage enriching, web services are not the best options. Web services are better used on smaller data sets, which are closer to the user. When data can be enriched with the external data without the need of transforming the existing data model, then it is advised to add the data in the earliest stage.

The next opportunity to add the external data to the system is integrating it at the core of the model. The paper mentions that this ensures the greatest level of adherence to enterprise-data

standards and data-management processes. Integrating the data at the core leads to clean data with the least mismatches. This is however the most costly option to maintain for an organization. Especially in organizations that work in rapidly changing environments. This is why it is not the most popular stage to implement the external data.

Another option is to insert the external data into a datamart. A datamart is a subset of the datawarehouse which has been designed for a specific purpose within the organization. Maintaining is easier in a datamart, because it is a smaller data set than the whole database. Organizations can use this if different departments, want to analyze different data. However, combining information from two datamarts later on if departments want to compare results has been shown to cause issues. This is why integrating external data in data marts is only advisable if the data mart does not have to be compared with other datamarts.

Most BI tools also provide options to combine internal and external data. This approach focuses on the end-user and requires no data modeling expertise. Within the BI tool is a set of option which can be combined with the internal data of the organization. This means that someone without access to the database structure and knowledge about modeling can also combine internal and external data. The external data can be easily added and directly be viewed, which is very user-friendly. However, the user is restricted by the options that the BI tool vendor provides.

The last stage to add external data is using the web scraping technique discussed earlier and adding this data to your database using a web service or using a simpler method to compare information like results. This method is the easiest to achieve but has the lowest accuracy and traceability. It can be used to see if there is a connection between multiple factors, but it is advised to use modeling techniques to find out how strong these relationships exactly are.

T. Löfgren, D. Gravem and G. Haraldsen have also performed research on the use of external data in their paper: A glimpse into the businesses' use of internal and external data sources in decision-making processes (Löfgren, Gravem, Haraldsen. 2011). They performed interviews as multiple Norwegian companies to find out what the use and potential was of NSI statistics. NSI statistics are statistics from the national statistical institute. This is a large database with information about its country ranging from education level of people to retail sale. This data is publicly available and can be used as external data source. The questions and answers that have been given during the interviews have been summarized in the following table.

Questions	Answers
Do businesses use NSI statistics?	Yes, and there seems to be a growing demand. The majority of direct request seem to come from professional services, transport, trade and catering industry. Who are the most active website users and readers of other kinds of publications and media coverage, however, is not known.
If so, what data?	Data that is relevant to marketing analyses and strategies. What is considered most relevant are thus data about prices, inflation and purchasing power. Next, more general demographic data that can be used to identify present and future interesting markets are also used. Finally business performance indicators can serve as productivity benchmarks. This kind of indicators is however generally not produced by NSIs.
What are the major obstacles preventing use of NSI statistics?	Lack of timeliness and relevant breakdowns are the main obstacles. Relevant NSI statistics may also be hard to find and the figures may be hard to interpretate. Businesses seek analyses and other processed information, not just data.
Who are the users of NSI statistics	There seem to exist an information gap between larger businesses, often with an international orientation, that are well trained in seeking information and making sense of figures, and smaller firms with less analytical resources.
How close is the relationship between data user and providers in businesses?	The larger, more professional and more active the businesses are when it comes to utilizing NSI statistics, the weaker their contact are with the data providers within their company seems to be.

Table 9: Research interview results, NSI statistics.

This table shows that companies do want to make use of external data but that they do not always find the time to do it and do not know what external data will be relevant for them. The lack of time for external data integrations shows that companies do not yet see the real value of external data otherwise they would focus more on this data and make time for it. Knowing which data to use from which data set is also an obstacle as shown in the interview results. The paper from T. Löfgren et al(2011) explains the model by Eurostat 2009 on quality of data. This model exists of a set of questions that an organization can answer before acquiring data to see if it is a good fit for the organizations. The model consist of the following focus points:

1. Relevance, which is the degree to which the data meets user needs both in coverage, content and detail.

2. Accuracy, which is the closeness between an estimated result and the (unknown) true value. Accuracy is a central part of the surveyors' professional approach to survey quality

3. Timeliness, which is the degree to which data produced are up to date.

4. Punctuality, which refers to the time lag between the actual delivery date of the data and the target date when the data should have been delivered.

5. Accessibility is the ease with which users are able to access the results, also reflecting the format(s) in which the data are available and the availability of supporting information.

6. Clarity refers to the quality and sufficiency of the data documentation, illustrations and additional advice provided.

7. Comparability, which is the degree to which data can be compared over time, spatial domains and sub-population.

8. Coherence, which is the degree to which data derived from different sources or methods, but which refer to the same phenomena, are similar.

These eight points however are towards data in general. From the interviews that have been conducted by T. Löfgren et al(2011), the quality dimensions from Eurostat rated by business representatives towards external data shows the points which should be taken into account for external data use. Which is: relevance, accuracy, timeliness and comparability

The managers of the organizations that were interviewed for the research found it most important if the data in the data set was usable for their own research. If this was the case they would look how accurate it was. This was not done by looking at the data specifically but by looking at the organization that published the data. If this organization was deemed objective, accurate and reliable then the data was inserted into their own database, as long as the data was up to date and comparable with their current dataset. What is important for the timeliness is that while it is important to have recent data, this recent data can only be used if the dataset is big enough. It was mentioned during the interviews that: "More recent information is of course more interesting, but perhaps less credible if there is a shorter time series behind".

To summarize, based on the literature an overview has been created on what external data is and when it can be implemented. There is no one best version of external data. It can come in different forms and an organization should look at its own preference to determine what external data should be used and wherein the process it should be implemented. There are however a set of aspects that are most important to keep in mind when choosing external data and these are: flexibility, quality and cost according to the first paper and relevance, accuracy, timeliness and comparability according

to the second paper. The aspects of the second paper all fall under quality of the data from the first paper, so this can be seen as a sub-set of focus points to complement the first paper.

4.4 Retroduction

In this sub-section the identified relationships between components of Van Eijck and the external data will be discussed. Like mentioned earlier, the tool used for the analysis is Power BI and the external data is a data set from both the KNMI and MSN on the weather. Relationships have been identified by performing interviews with Van Eijck, getting a tour through the entire database and exploring the database with the help of visualizations and sub-columns created using the DAX, the coding language in Power BI. The findings have afterwards been discussed with experts from Lumen to determine the elements on which to focus, based on their possible contribution to effective changes within Van Eijck. The relationships identified by the exploration are mostly macro-micro mechanisms. This means that a large subset of data influences and constraints various parts of the organization. The large data subset is the weather data. The interaction between the weather data and part of the system of Van Eijck is what is a mechanism. The mechanisms in this section show the influence of the temperature, windspeed, visibility and humidity and the number of incidents on the road, the arrival time and the distribution of services. The database from which all the visualizations are made can be seen in figure 10.

4.4.1 Arrival times

To find the relationship between the different key aspects, visualizations had to be created to show how different key elements react to each other. Like mentioned earlier in section 4.2.2. the main interest of Van Eijck is keeping the rayons that they have been appointed to and acquire new ones if they can handle that. To keep and get the rayons the time that they take to arrive at an incident should be as low as possible. That is why the first focus point was visualizing how the different rayons are performing. 4.2.2 also mentioned that salvage companies should arrive on time at the incident before a benchmark set by the government. This 20 minutes during daytime and 25 minutes during night-time. In addition the salvage company has to arrive at incident before a benchmark that they present themselves. This means that an employee of a salvage company can arrive at the incident before the benchmark of the government but still be too late due to the benchmark of the salvage company itself. Van Eijck wants to know how they perform both on the benchmarks that they proposed and the benchmarks from the government. This is important to know, because it will show which rayons Van Eijck is at risk of losing and should be focussed on to improve.

To do this the data had to be split into incidents that happened between 23:00-05:00 and incidents between 05:00-23:00. Next was looking at the arrival time of the employee to see if they were on time at the incident. New tables were created for this process that looked the following:

Hourofexecution	DagofAvond	Voldaan
11:50:00	Overdag	Optijd
11:59:53	Overdag	Optijd
15:31:54	Overdag	Optijd
10:59:29	Overdag	Optijd
14:26:27	Overdag	Optijd
06:00:00	Overdag	Optijd
07:00:10	Overdag	Optijd
13:02:51	Overdag	Optijd

Figure 12: Determine incident time

Using this distinction between day and night incident, calculations could be performed to see if the employee on time. If a rayon is healthy, then 90% of the time employees of the Salvage company arrive on time. If they arrive 80% or more on time, then it means that the rayon should be investigated to see how it can be improved but it still has an acceptable score. If the incident arrival is below 80% then the situation of the rayon is critical and it should be investigated whether the rayon can be kept by improving the stations near and in the rayon, the benchmark time of the salvage company can be increased or that they have to conclude that they cannot handle the rayon and should let it go. The status of rayons can be evaluated with the following visualization that has been created for the research.

Figure 13: visualization of incident arrival time

Figure 10 shows how a few of the rayons are performing based on the time it takes to arrive at an incident. The block with "% vaste waarde optijd" shows how the rayon is performing with regards to the benchmarks set by the government, so the 20/25 minutes benchmark. The block with "% contract optijd" shows how the rayon performs in regards to the benchmarks set by the salvage company. The background color of the rayons is the most important focus point for this visualization. It can be one of three colors: green, orange or red. If both benchmark times are above 90 percent the background is green and the rayon is performing well. If the background is yellow then it means that one of the benchmarks is between 80 and 90 percent and should be monitored but is not concerningly low. If the background of rayon is red then one or both benchmarks are below 80 percent which leads to a high reason for concern. This visualization shows which rayons are underperforming but do not give an explanation for the bad arrival times. To find this out another method should be used.

4.4.2 Weather data

For this research and interest of the salvage company the correlation between the weather and the arrival time of incidents at an incident, but also the number of incidents that happen on the roads of each rayon per hour are interesting. Because the historical data only has information about the windspeed, visibility and humidity, the focus has been mostly on these aspects of the weather. To see the effects of these weather aspects, it was looked at how many incidents happened within an hour at certain weather states. This lead to a number of visualizations. These first visualizations have

been made based on the interviews with Van Eijck and exploration though the database. The results will show an indication of the effect of certain weather aspects and the incidents per hour. Later on in the research the significance will be tested of the visualizations which show interesting results.

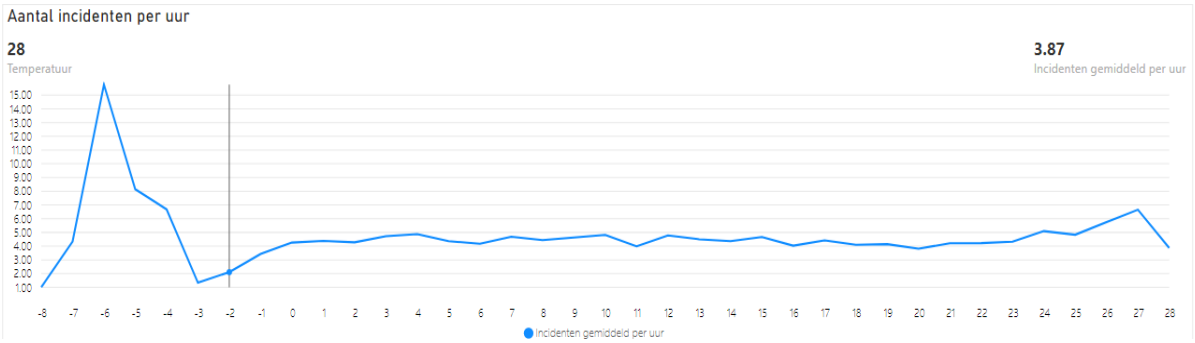


Figure 14: Average incidents per hour compared to temperature

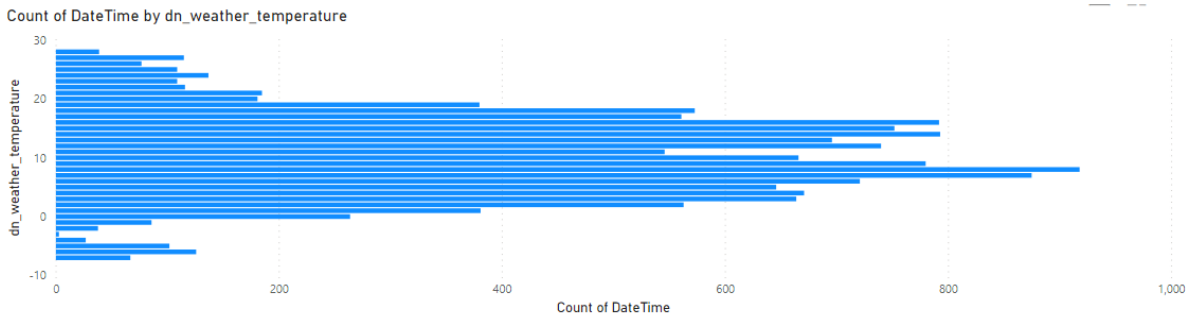


Figure 15 Amount of hours that temperature occurred

The first external aspect analysis inspected how the cold affected the number of incidents that happened on average within an hour. The graph clearly shows the peak when the temperature drops below minus three degrees furthermore the line is relatively straight. The average number of incidents per hour is normally around four, but when temperature drops below minus three it rises all the way up to 15 incidents per hour. Since the graph shows average incidents per hour it can be concluded that on average the most incidents happen around minus six degrees. Figure 15 shows the number of hours that it was a certain temperature, for this graph it was decided to discard the data from the temperatures that occurred less than 100 hours. This means that based on figure 15, nothing significant can be said about the results of the temperatures -1, -2, -3, -4 and -7. This means that the increase in incidents could start before minus five but more data should be acquired to determine this. Another important aspect for later research is if the roads have been sprinkled with salt or not. This information could also be included in further research to give more meaning to the data.

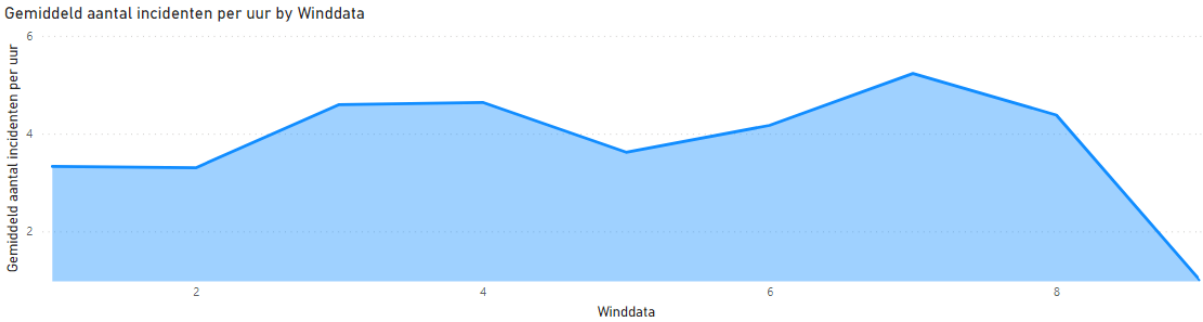


Figure 16: Average incidents per hour compared with windspeed

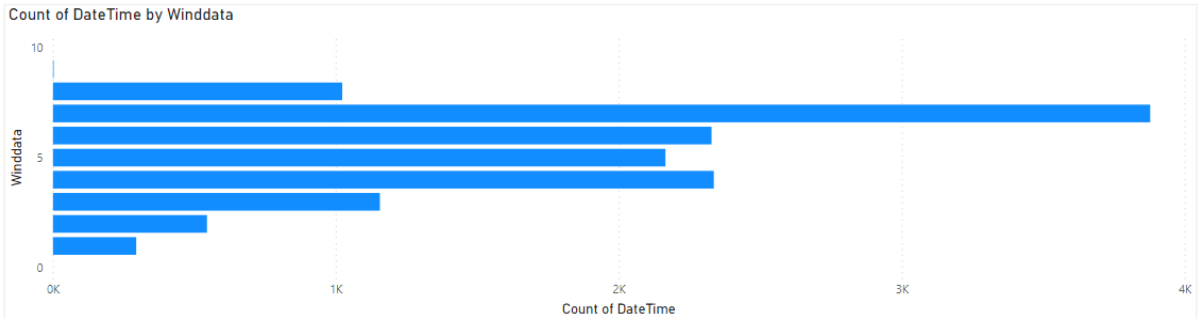


Figure 17 Amount of hours of each windspeed

The data from figure 16 shows that the average amount of incidents increases slowly when windspeed increases and surprisingly decreases a bit when windspeed 8 is reached. The most dangerous windspeed is 7, since most incidents per hour happen on average at that windspeed. Figure 17 shows the number of hours it was a certain windspeed. For the windspeed it was determined that has a lower spread than temperature, a windspeed should occur at least 500 hours for it to produce usable data. The data from windspeed 1, 9 and 10 do not give an accurate representation of reality since they have occurred less than 500 hours. This means for figure 16 that while the peak at windspeed 7 and the slight decrease at windspeed 8 are a good representation of reality, the heavy decrease at windspeed 9 and higher are incorrect and more instances of windspeed 9 or higher should be added to the database to perform an accurate evaluation on these windspeeds.

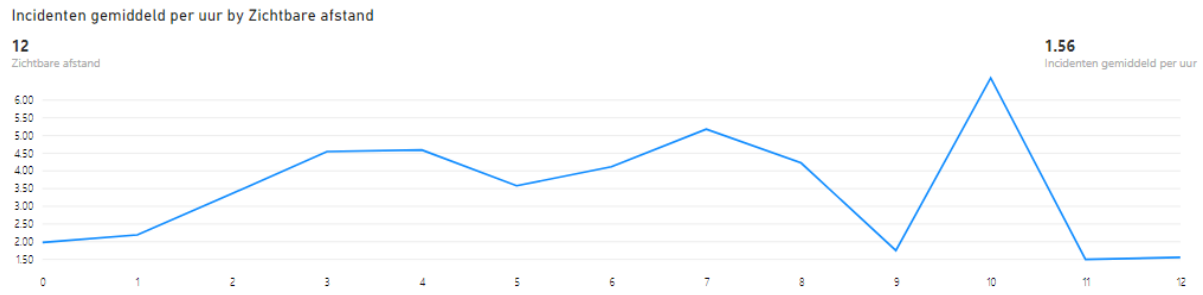


Figure 18: Average incidents per hour compared with visibility on the road

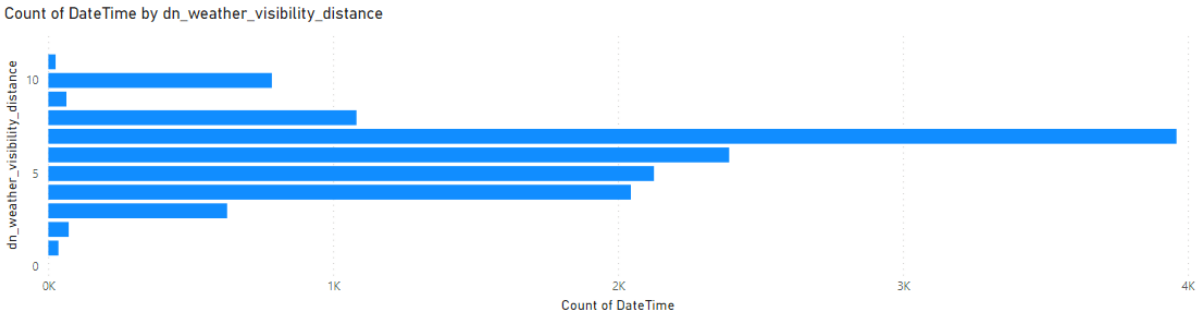


Figure 19 Amount of hours each visibility occurred

When it gets foggier outside and the field of vision of drivers decreases then it gets more dangerous on the road is what is expected. This is however not shown in figure 18. The opposite even, the number of incidents per hour is the highest when it is clear outside. Here it could again be the case that more cars get on the road on clearer days or drivers are blinded by the sun on clearer days. Figure 19 does however show that visibility below 4 km does not happen more than Further than 500 hours. This means that no significant conclusions can be drawn for the instances that the visibility is bad. More data should be acquired to determine the effect of low visibilty on the average amount of incidents per hour.

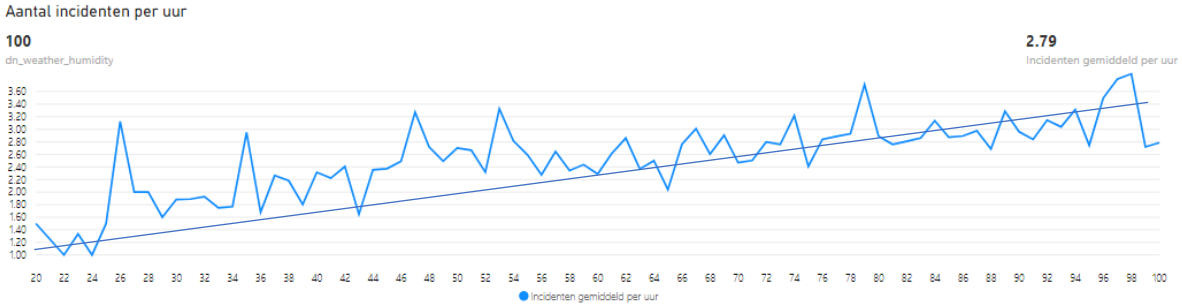


Figure 20: Average incidents per hour compared to the humidity

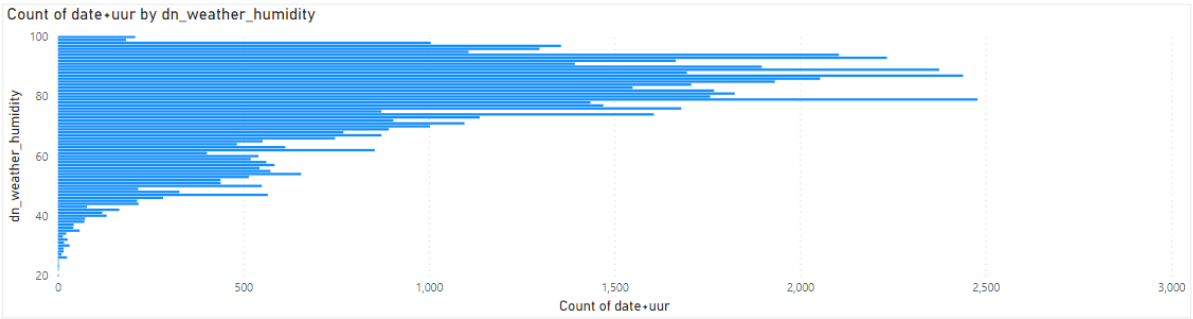


Figure 21 Amount of hours each humidity % occurred

Figure 20 shows how the average incidents per hour are affected by the humidity. At first it was the plan to compare the number of incidents to the rainfall, but the historical data that has been added to the database does not show this data. That is why we have gone for the humidity. A higher humidity does not necessarily mean that it is raining, but it does mean that there is a higher chance that is raining or has rained. With that in mind it can be seen in figure 20 that if the line is plotted, a correlation between the humidity and incidents per hour can be spotted. For this visualization applies that this is a general picture of all rayons and a closer inspection should be performed to see which rayons are influenced the most. However figure 21 shows that a humidity below 47% and

above 98% have not occurred enough hours(each percentage less than 500 hours) to give significant conclusion on these humidity percentages in further analysis.

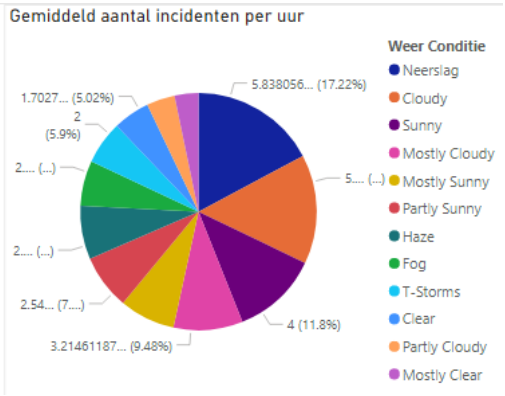


Figure 22: Average incidents per hour compared with weather conditions

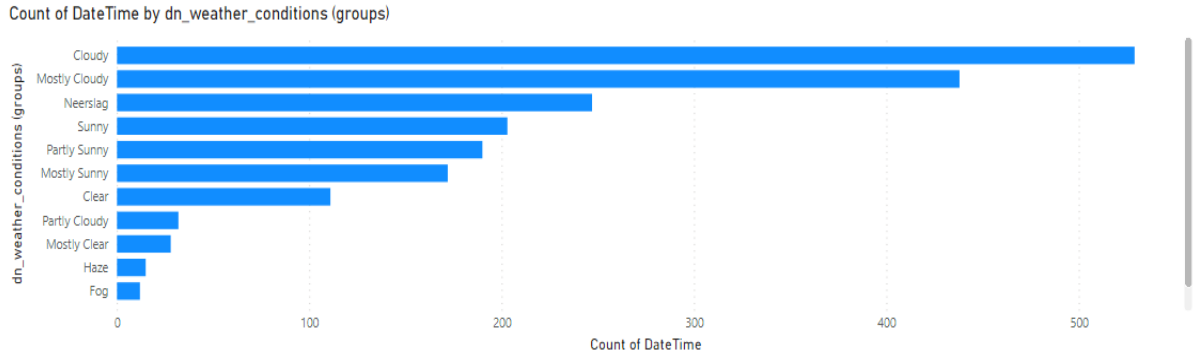


Figure 23 Amount of hours weather conditions occurred

Like mentioned earlier, the historical data added to the database has no weather conditions. However, the data that was inserted since the start of April do have weather conditions. The number of incidents that have weather conditions attached to them is currently 9493 entries and increasing every day. The data that is generated by this subset might not be as reliable as the entire dataset but still gives an idea of the correlation between the weather conditions and the incidents that happen per hour. The MSN database has a large set of weather conditions, but for this research the condition “Neerslag” has been added which is a combination of the following conditions:

Neerslag

- Light Rain
- Light Rain and Snow
- Light Snow
- Rain
- Rain Showers
- Snow

This combination of conditions shows the amount of incidents per hour for every type of rainfall from MSN weather. This visualization tells that when the weather condition is rainfall, the average incidents per hour is 5,8. Cloudy is the second largest part of the pie and has 5,06 incidents per hour. This means that during rainfall almost one incident more happens per hour. This confirms the findings from figure 20 about the correlation between the humidity and the incidents per hour. Salvage stations should have more cars on standby whenever rain is predicted to make sure that the

arrival time does not get too high with the heavier workload. For the weather conditions all the conditions below a clear condition in figure 23 are not usable for the research, since they have occurred less than 100 hours in total thus far. The database should grow to give significant conclusions on these conditions. Rainfall does occur enough times to say that significantly more incidents happen on average in an hour during this weather condition.

To create a better overview of the influence of the weather on the number of incidents that happen on the road per hour three categories have been created. These are: good weather conditions, mediocre weather conditions and bad weather conditions. Good weather conditions include all the incidents where the weather was sunny and clear and there were low windspeeds. Mediocre weather condition are applicable for all the incidents where the weather was cloudy or light rain with only a bit of sun and wind speeds of around 10 km per hour, which is considered mediocre by the KNMI. The last weather condition is bad weather, this includes the incidents where the windspeeds were above 15 km per hour, there was rain, haze or snow and it was cloudy. This resulted in the following visualization.

Figure 24: The average amount of incidents per hour for weather groups

Figure 24 represents the number of incidents during bad weather on the left side, in the middle the mediocre weather conditions and at the right the good weather condition. Noticeable from this visualization is that the amount of incidents during mediocre and good weather is almost the same, the amount of incidents during good weather is even slightly higher than the mediocre one. This could again be due to more people going on the road to visit places during nice weather and the possibility that people get the sun in their eyes when the brightness is high. In future research the direction the car was headed can be added to the database, because this combined with the amount of UV can be used to determine if an incident might have happened because the driver got the sun in their eyes. The bad weather shows a significantly higher amount of incidents that happen per hour, 5.78. This is 1.5 more than the average of the other two. This indicates that if the conditions of the bad weather conditions are predicted, more salvage cars should be ready on standby, because a heavier workload can be expected.

So far the analysis has been performed but to get more meaningful conclusions the situations of separate rayons should be investigated. Every incident has a latitude and longitude, using these coordinates the incidents can be tracked back on the map. This made it possible to show the incidents on a map and include filters to see where, with certain weather conditions the most incidents happened.

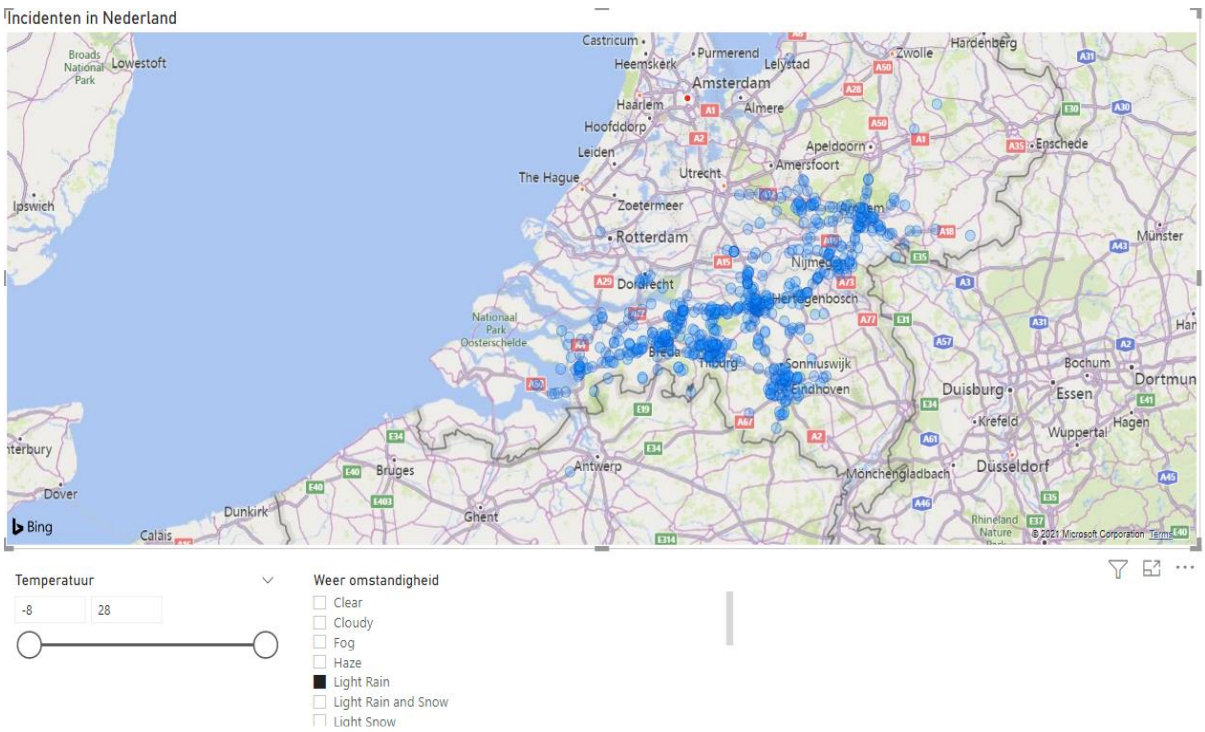


Figure 25: Incidents on the map during rain at rayons of Van Eijck

Figure 25, shows all the incidents that have happened with the conditions rain and light rain for Van Eijck. It is noticeable that some areas are darker blue because more incidents have happened there during the rain. When taking a closer look at these incidents it is found that the darker blue areas are near big cities like Eindhoven, Breda, Tilburg and s-Hertogenbosch. It is logical that near big cities more car incidents happen, because the car density is higher at those places. Using this map it cannot be examined if the number of incidents increase at certain places because of the weather. This is why other visualizations are needed to inspect the real influence of the weather conditions at the roads in the rayons that are owned by Van Eijck.

4.4.3 key influencers

In the previous section, graphs were shown which were built based on information provided by Van Eijck and information which had been found in the database during experimentations with the data. These graphs have shown the first effects of the weather on the incidents on the road, but they do not show how significant the significance is. This is why the findings from section 4.4.2 have been analyzed using a function from power BI which shows only relationships when they are significant. This is important for the analysis because conclusions should not be drawn from random occurrences.

The function in Power Bi to use when investigating multiple factors from which you do not exactly know how they influence each other is the key influencer function. This visualisation makes it possible to select an element of the database that you want to inspect and after this you can insert multiple other elements from which you suspect that they influence the main element. This visualization takes into consideration the number of datapoints that are used if the n (number of data points) is not high enough then the element will not be appointed as a key influencer. Besides the size of the n, the key influencer function also looks at the p-value. The p-value shows the chance that the relationship between multiple elements has occurred by random chance. The p-value of a

relationship in the visualization should be below 0,05 to be valid. This means that the chance that the relationship between the elements is random is lower than five percent.

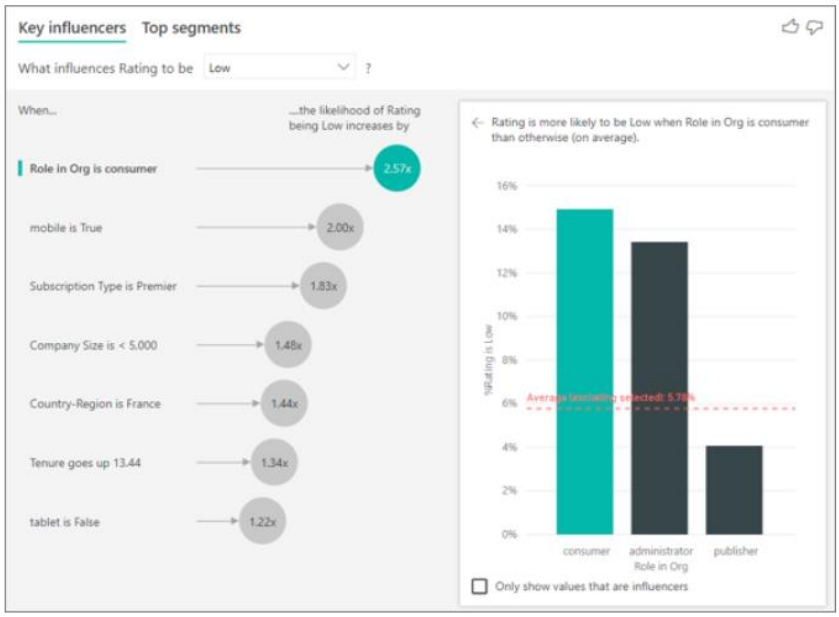


Figure 26: Example of key influencers

Figure 26 is an example of the key influencer function in a Power BI dashboard at another company. This function will also be used on data from Van Eijck, based on the previous visualization the elements that will be tested using the key influencer function are how the temperature, visibility, humidity and weather conditions influence the amount of time that an employee takes to arrive at an incident. Based on the previous analysis it is expected that humidity has a positive relationship with the arrival time and temperature has a negative relationship with the arrival time. Using a filter on the rayons, removing the incidents which did not occur on an IM road, these are the roads where incident management is active and removing the incidents that were discarded later on by employees the rayons were found that were most influenced by the weather. Rayon names are not shown, to keep the information for Van Eijck.

4.4.3.1 Exceptions in four rayons

At most rayons, no influencers were found using the current set of influencers, which means that the arrival time of employees to the incident is not influenced by the weather. Four Rayons did however have significantly different arrival times when the weather changed. These will be discussed here as rayon A,B,C and D.

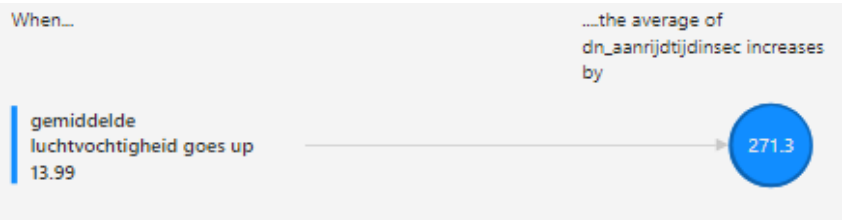


Figure 27: Results rayon A

At rayon A the correlation between the arrival time and the weather was that if the humidity increases the average time an employee needed to arrive at the incident increased by 271 seconds. This is quite substantial. When we take a closer look at the information this is mostly because the

arrival time at some incidents is 0 seconds. A 0 time can be achieved if the salvage company figures out that an incident has happened somewhere before they are officially notified.

Figure 28: Results rayon B

Rayon B shows more results. This rayon is not only influenced by the humidity going up but also by the temperature. As expected when the temperature drops the arrival time at this rayon increases, in this case the arrival time increase by more than ten minutes for both the temperature and the humidity. An increase of 10 minutes lowers the chance that employees will be on time at the incidents extremely. This rayon should be investigated to determine why arrival times increase this much when the weather gets worse.

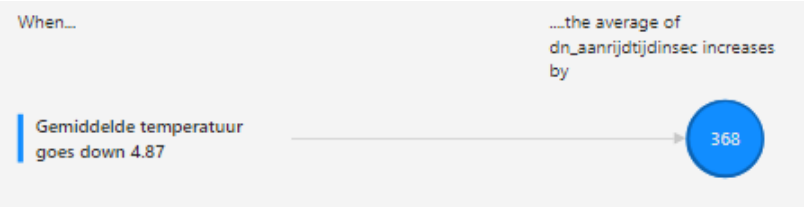


Figure 29: Results of rayon C

Rayon C is also interesting to look at with the weather data. Just like rayon B, the arrival time increases by more than ten minutes if the temperature decreases. When we look at the average time there are no zero cases like with rayon A. This means that the stations near rayon C should be investigated to find out what causes this increase in arrival time.

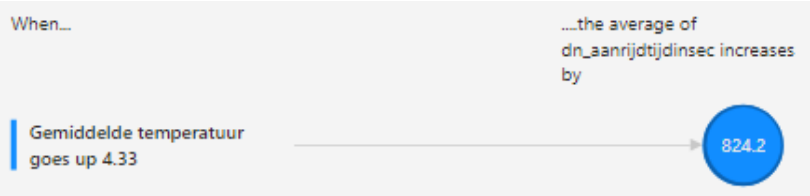


Figure 30: Results of rayon D

The last rayon with significant data is rayon D. Here the opposite of what is expected happens. When the temperature increases, the time it takes to arrive at an incident also increases. This increase in time is also extremely high, more than 20 minutes. When we take a closer look at the data it is found that there is one incident with an extremely high arrival time which is above 6000 seconds. This is an error in the data and if this data is removed from the table, there are no significant relationships in rayon D to examine.

4.4.4 Service/weather influence

In section 4.2.3 the different kinds of services are discussed which a salvage company provides. One of these services is *roadside assistance*. This type of assistance is needed when a car breaks down and the salvage company has to come to repair the car, with the hope that the car does not have to be towed. It is expected by Van Eijck that this type of service is highly affected by the temperature. This will be investigated in the next visualization.

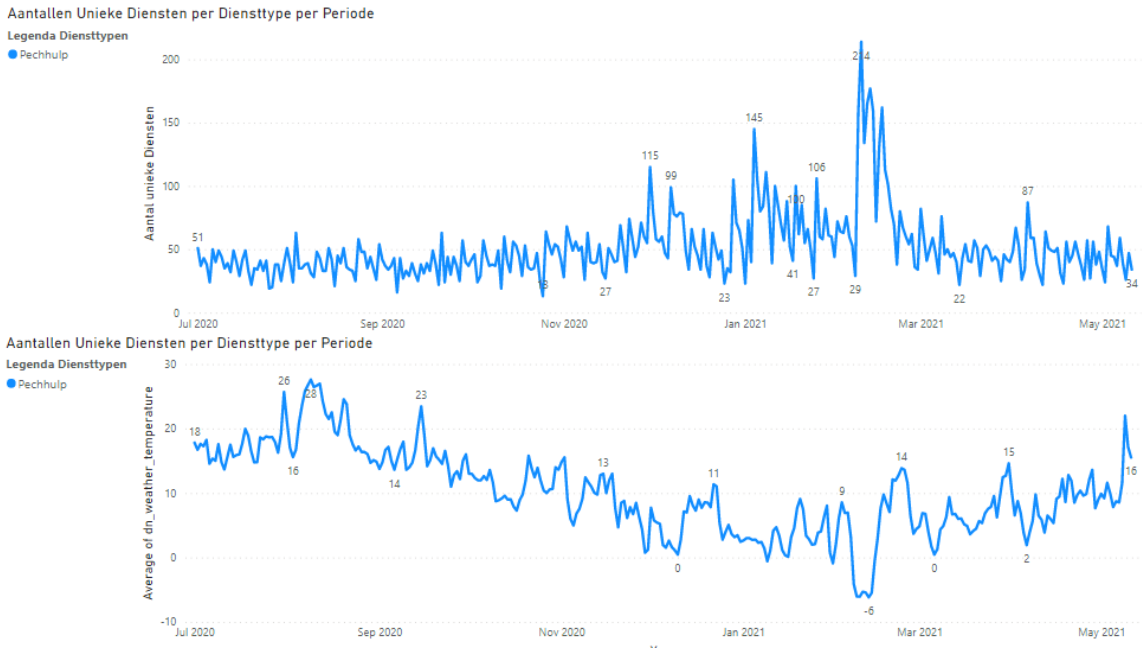


Figure 31: Temperature compared with roadside assistance

The expectation is confirmed by the valuations. Especially in February 2021, the data shows the correlation between roadside assistance and the temperature. When the temperature reaches its lowest point of minus six degrees, the amount of road assistance hit a peak. This means that when low temperatures are expected the salvage company has to be prepared to perform more road assistance. Within road assistance there are multiple types of assistance. If employees of the salvage company know what type of assistance they can expect on colder days they can be better prepared and equipment can be placed in more accessible places or more equipment can be put ready on colder days.

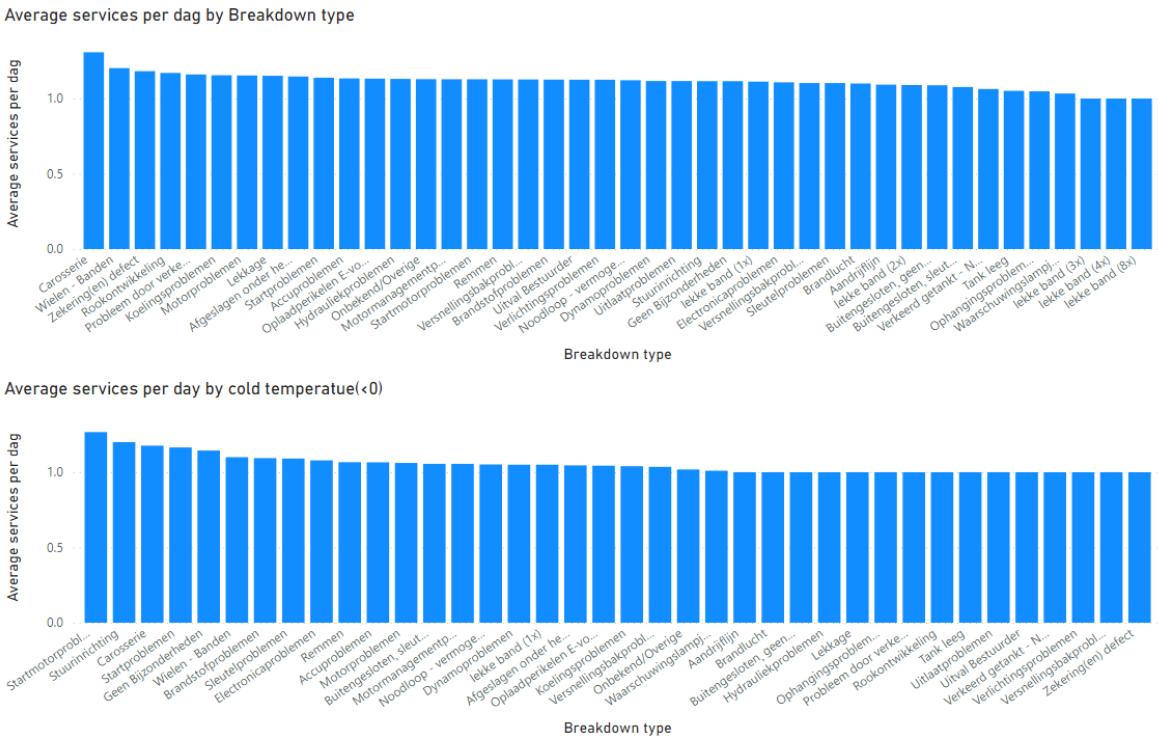


Figure 32: Breakdown types compared with the temperature

Figure 32 shows what the most popular types of breakdowns are for road assistance in general and when the temperature drops below zero degrees. This visualization shows the average amount that a kind of breakdown assistance is requested. The amount of requests for each type of breakdown is quite close, but a small difference is still noticeable. During temperatures below zero breakdown type start problems are the most requested on a daily basis while it is normally corrosion in the car. Another type of breakdown that is more popular during negative temperatures is problems with the steering gear. The rest of the breakdown types seem the same as normal. For the correlation between the temperature and the type of service a key influencer analysis is again needed to determine if during the cold there is indeed a higher amount of roadside assistance compared to the other services than normal.

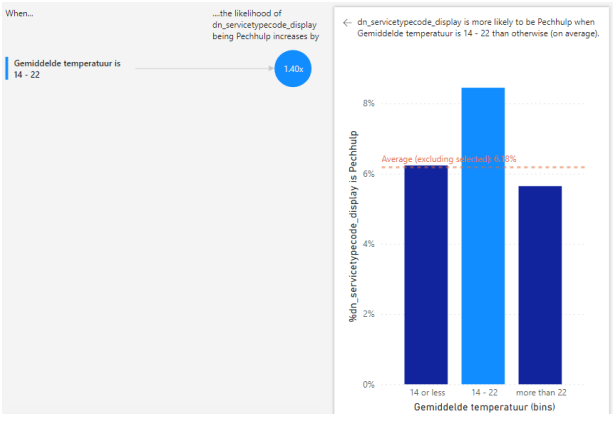


Figure 33: Key influence temperature/ roadside assistance

According to the analysis, the amount of roadside assistance is the highest compared to the other type of services when the temperature is between 14 and 22 degrees. This means that although

there is a higher amount of roadside assistance calls when the temperature is below zero degrees, all the service types increase in calls and the roadside assistance not significantly more than the rest.

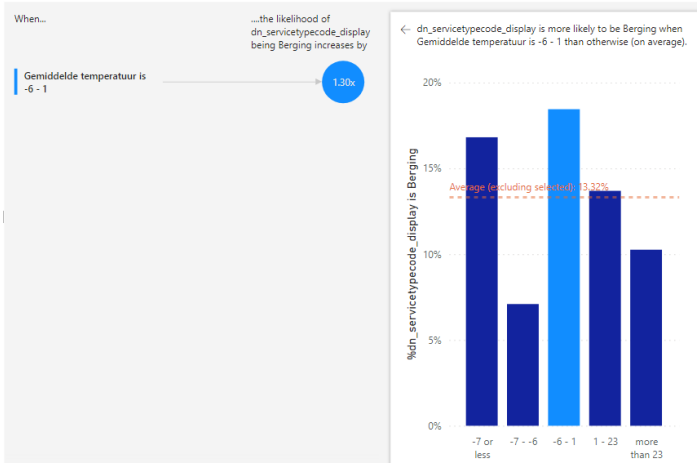


Figure 34: Key influence temperature/salvage

All the other service types have been analyzed too and the only service type that has a significant increase is the salvage like shown in figure 34. When the temperature drops below zero degrees the number of incidents with type *salvage* increases times 1.3 as can be seen in the figure. For heavy salvage it is even higher, but with heavy salvage the increase starts at five degrees. Below five degrees the number of heavy salvages increases with 1.5 and below minus five degrees it even doubles compared to the average state of 4%. This means that although road assistance was expected to increase the most with colder temperatures, the actual truth is that salvage and heavy salvages increase the most when temperature drop below zero degrees.

4.4.5 Windspeed effect on rayons

The next focus point is how the weather influences the amount of incidents that happen in the rayons. In an interview with Van Eijck to discuss current findings, they said that they expected that certain rayons would have more incidents when the windspeed was high so this was investigated.

To investigate the possible relationship a new dashboard was created that shows the number of incidents that have happened in all the rayons and a slider to increase or lower the wind speed.

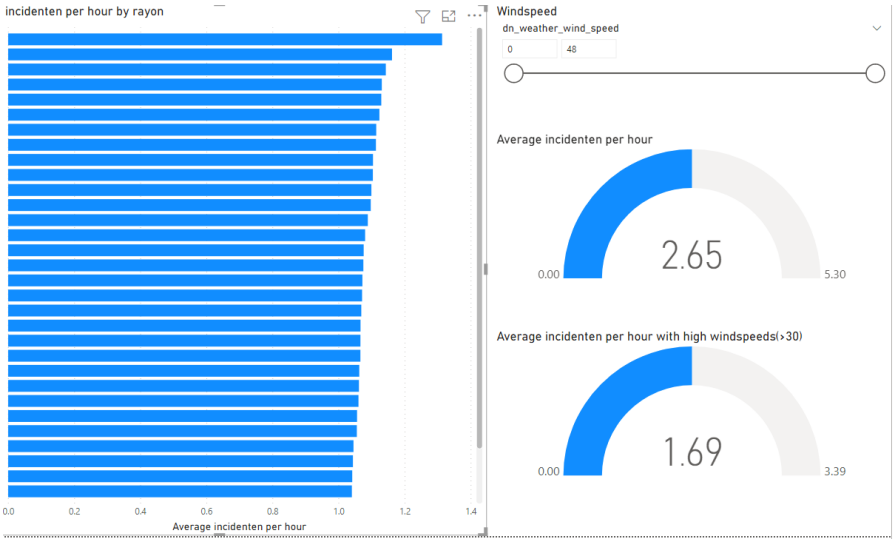


Figure 35: Windspeed/rayons dashboard

At the right side the average incidents and the average incidents that happen when wind speed is above 30 km per hour are shown. Different rayons can be selected to see how it changes the averages. For this figure, the rayons are again cut off the picture to keep the data hidden for competition. According to the KNMI windspeed, five starts at 30 km per hour and this translates to a quite powerful windspeed. For this reason 30 km per hour and everything above it has been selected as a benchmark for a strong wind speed. Figure 30 shows that when all the rayons are selected the average incidents per hour is higher if all the windspeeds are taken into account. This changes however when the slider is increased to 30 km per hour. When this happens two new rayons become the rayons with the highest average of incidents per hour. These new number one and two will be called rayon A and B for the analysis. Rayon A shows the following results when inspected more closely.

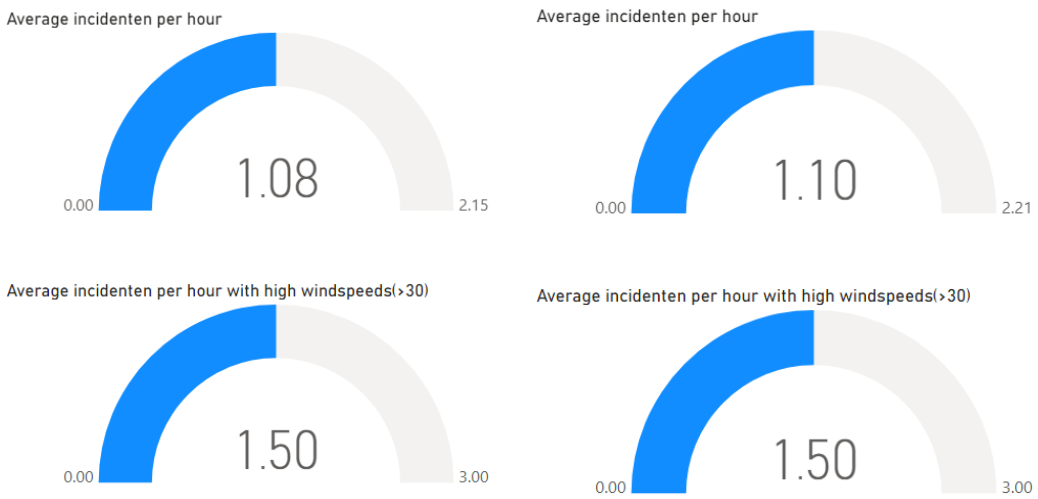


Figure 36: Rayon A & B windspeed effect

Rayon A shows the biggest increase in the number of incidents that happen per hour. Per two hours one more incidents happens in this rayon if there are windspeeds above 30 km per hour. More salvage cars should be on standby if high windspeeds are expected, although the actual significance will still be tested to determine if this increase is not random.

Rayon B shows similar results as rayon A. Here the amount of incidents per hour is also almost 50% higher if the wind speeds are above 30 km per hour. This rayon will also be examined further for the significance level. The analysis for the significance level will once again be performed using the key influencers combines with filters for the specific rayons. Besides the influence of wind speeds, which have been investigated in the previous visualization, the effect of the temperature, humidity and weather conditions will be tested to see the effect of these aspects on the amount of incidents that happen within a rayon. Starting with the rayons that showed to have the biggest relationship with the wind speed.

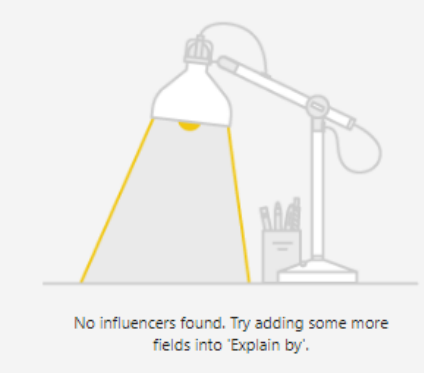


Figure 37: Effect windspeed on rayon A & B

Unlike what was expected the windspeed has no significant relationship with the amount of incidents that happen within an hour. Rayon A & B might still be investigated to see why they have higher peaks when there is a strong wind. The strong wind might often be combined with more rain which could also lead to more incidents. These suspicions will be shared with Van Eijck and the employees that drive in rayons A & B. They might have an explanation based on their experiences on the road what causes rayon A & B to perform worse but not significantly worse when there are strong windspeeds. During the significance research there was one interesting finding.

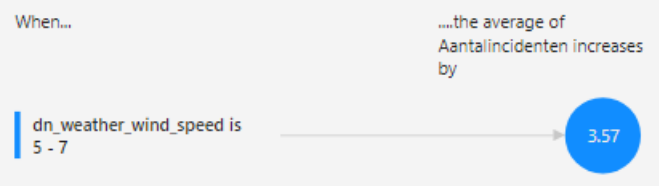


Figure 38: Effect windspeeds on accidents per day

The average amount of incidents during an hour does not increase but the average of incidents during the whole day does indeed increase. The Key influencer shows that if the wind speed is between windspeed 5 and 7, which translates to winds that are quite strong to extremely strong, the number of incidents that happen on a day increase by almost 3,5. This increase is most likely in rayon A & B, but the increase in these rayons is not significant enough to come forth from the key influencer analysis.

4.5 Analysis of mechanisms and outcomes

In this subsection the causes of the relationships between the different elements will be discussed. These are the relationship between the internal data elements and the internal/external relationships.

Like discussed earlier, the most important aspect of the data for Van Eijck at the moment is the arrival times at accidents. This time should be reduced to keep contracts. Noticeable from the data is that the arrival times at incidents have a lot of room to improve shown in figure 12. The external weather data has shown that certain weather conditions increase the arrival time at an incident or the number of incidents on the road. Experience has learned us that in case there are more incidents on the road, that more employees arrive too late at the incident. There is a restricted amount of vehicles available to send on the road, so more incidents mean less availability. To increase the availability the arrival time and time it takes to handle the accident should be reduced.

There are multiple relationships between the amount of incidents per day and the weather. Figure 14 explained is the amount of incidents that happen compared with the temperature. It shows that

when it gets colder more incidents happen. This can be explained by the road getting more slippery during temperatures below zero degrees. Below zero degrees rain turns into snow and hail which makes the roads more dangerous. This is the logical explanation for the relation but can now be shown using the visualization. Minus six degrees seems to be the point that the most incidents happen, most likely because the roads are dangerous but people are still willing to go on them apparently. After minus six degrees people avoid the roads and the number of incidents reduce.

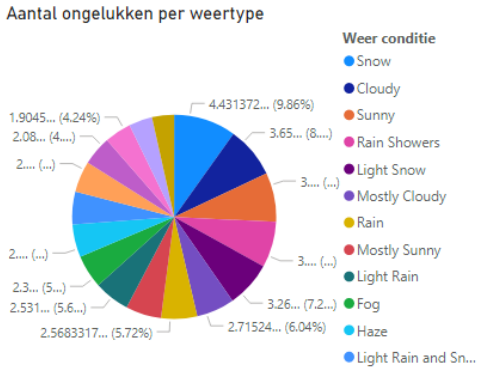


Figure 39: Average incidents per hour without clustering of weather conditions.

Figure 33 further confirms the influence of cold weather on the average amount of incidents during one hour. When it is snowing almost one more incidents happens per hour than during the cloudy weather conditions. The snow is significantly more dangerous than the haze, this is because the snow makes the road more slippery than the haze.

When the windspeed reaches above windspeed 6 the amount of incidents per hour increase can be seen in figure 16. This measurement is taken from highway road. Cars are highly effected by gust winds due to the Netherlands being flat. The chance is higher that people lose control over the steering wheel when the windspeed is high, which causes there to be more incidents.

The humidity in figure 20 shows also a positive correlation with the amount of incidents on the road. A high humidity means that it has most likely rained or is raining. When humidity reaches above 82% the amount of incidents per hour get above average. This can be explained by the fact that rain makes driving more dangerous because the roads get more slippery and the vision is decreased by the rain falling on the windshield. The danger of rain is further confirmed by figure 18. This figure shows that the any type of rainfall increases the amount of incidents per hour by 0.5. Which is a lot knowing that less than four incidents per hour normally happen when it is cloudy, the second largest weather condition when it comes to incidents per hour.

The increase in arrival time at the incidents in rayon A, B, C and D discussed 4.4.3. is a logical consequence of the increase of incidents explained by the bad weather type above. Interesting is that these rayons seem to be more significantly influenced than other rayons by the bad weather conditions. Here the great influencers are the temperature and humidity. This can be explained by the location of the roads and their position, this will be further discussed with Van Eijck but the exact rayons cannot be named in this Paper.

The same applies to the two rayons from figure 36 which are more influenced by high windspeeds. The explanation for these rayons could be that they have more open areas than the other rayons which causes the cars to be more sensitive for the wind. Like the previous rayons, the exact location of the rayon cannot be discussed.

The relationship between the different type of breakdown services compared with the temperature in figure 31 and 32 resulted in the finding that during temperatures below zero degrees engine start problems and steering problems occur more often than normal. This can be explained by the effects of cold temperatures on a car. The cold temperature reduces the capacity of the battery making it more difficult to start. Old batteries are often victim to the cold temperature. The cold temperature can also have an effect on the power steering.

4.6 Validation of explanatory power

During the previous phases multiple relationships and their causes have been found. Not every negative relationship is significant enough to be fixed on the short term. Some relationships have the highest priority and their possible solutions will be discussed here.

The best start is by looking at the specific rayons where a negative influence of the weather was found. These have the highest priority because although the analysis has shown that overall the amount of incidents are influenced by the temperature, rain and windspeed. These rayons are more influenced than others proven by a significance test using the key elements visualization. The first case shown in figures 27 to 30 shows how employees of Van Eijck arrive slower at the incidents when the temperature drops or it rains more. This increase in arrival time at an incident should be lowered to ensure that no rayons are lost due to arriving too slow too often. The advice is to speak to the employees at these rayons to find out why they think they arrived later than expected during rain and cold temperatures. This conversation can motivate them to improve their work the next time they work slower than normal. If the problem is out of their reach, but due to a lack of available cars, make sure to place more cars at the station A,B,C and D when the weather conditions go up or down in accordance with the analysis of figures 27 to 30.

The other point of focus should be the two rayons in which there happen significant more incidents according to figure 36. These two rayons should have more incident cars available when cold weather is expected since these rayons are affected the most by the strong wind. Having more cars available is a good way to make sure that for every incident a car is ready when needed. This will avoid the situation that the arrival time of an employee to an incident is high because he was still working at another incident.

If the rayons which have been specifically selected through the analysis have been inspected and changed if possible then Van Eijck could look at figure 24 to change the occupation of all the stations that Van Eijck has though out the south of the Netherlands. The data shows that if bad weather is expected, which is indicated by cold temperatures, rainfall and strong winds, the amount of incidents increases by two. This means that more cars should be available and the number of people scheduled on these days should also increase. When the temperature drops below zero degrees, the number of heavy salvage services needed increases more than all other services. This means that more cars capable of heavy salvage and the people that can drive them should be available on cold days according to figure 34. For the other cars it is important to make sure that the tools needed for starting problems and steering difficulties are available in the truck when cold weather is expected because these types of services are more common during cold days.

The final advice is to let the database grow and see how it affects the current visualizations that have been made. The CRM database of Van Eijck consists out of data starting from April 2020. If more data is gathered over the next year, the conclusion from the data can become more valid. Currently 88,748 incident reports are in the database but this number can become a lot bigger. Because the data was only from roughly one year peaks in the data have a higher chance of being incidental. Once the database consists out of data from multiple years, the outliners can be better analyzed as

being real outliners or a pattern can be discovered in the data, that can be used to improve Van Eijck.

4.7 Validation of design process

To see how valid and usable the analysis of the data from Van Eijck with the external weather data is, Van Eijck has looked at the visualizations and results. Though out the project a close contact was kept with Van Eijck to make sure the data was not interpreted the wrong way and to discuss what interesting points were to analyze during the research.

During the meetings, certain filters had to be adjusted to make sure the data showed the right elements. Some services were no longer provided and had to be scrapped from the research and there were a few codes attached to incidents which meant that they were duplicates. These incidents had to be removed. Other incidents had a code that Van Eijck had got a call to go to an incident but that they were not really needed. These incidents also had to be removed from the analysis. When inspecting the external data closely we found that the wind speed was measured in different ways by MSN and the KNMI. The KNMI made use of wind power and MSN weather gave the windspeeds in kilometer per hour. This has been adjusted and now all-weather elements have the same measurement types when comparing data from the KNMI and MSN weather.

After the corrections advised by Van Eijck were implemented, they were happy with the visualization that had been made and the analysis that was performed. The effect of the cold temperatures on their services was especially helpful to explain anomalies in the data. They will also further investigate the rayons that were analyzed to be the most affected by the wind.

The dashboard created in figure 13 is also something that Van Eijck will continuously keep using. This dashboard showed whether incidents were on arriving on time at incidents on contract and regulation bases. They were pleased with this dashboard because it shows the user in an easy way which rayons are underperforming and should be analyzed further to avoid losing the rayon.

Since the data from the database is only roughly one year old, Van Eijck also wanted to wait longer to make serious changes in their organization based on the data. The plan at the moment is to gather three years of information and investigate the visualizations then again which have now been made. This way they can be sure that they do not change processes in the organization based on incidental occurrences. Figure 15, 17, 19, 21 and 23 are most important to inspect after the three years to find out if the different types of weather conditions have occurred enough to perform changes within the organization based on the weather data.

This means that the data analysis of the CRM data of Van Eijck combined with the external weather data was a success. The visualizations made will help them further in their growth and small changes can be made based on the recommendations in section 4.6. However, big changes will not be made at the moment. The visualization will be reviewed and analyzed again in three years and based on the results of this analysis big changes can be made in the organization.

Chapter 5 approach CRM analysis external data

Based on the findings from the literature, the interviews and the empirical data from the case study a first iteration of an external data integration approach has been developed to answer RQ6 (Kerckhoffs, 2021). Organizations can use it if they want to prepare their database for integration of external data and how they can get the most value out of this external data. This approach looks the following.

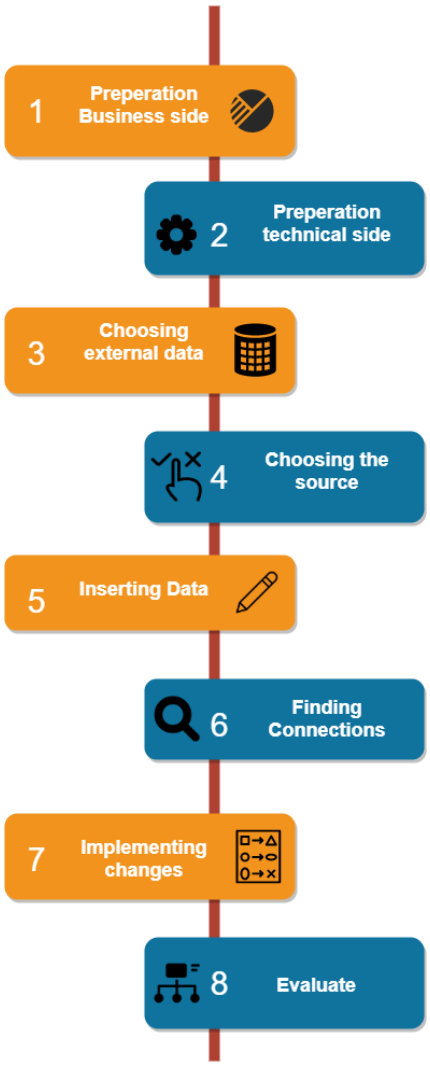


Figure 40: External data integration approach

By following the eight steps of the approach organization can more easily and in a structured way implement external data into the data base and make adjustments to their company based in the conclusions from the analysis. We will go in depth for every step to elaborate on the approach.

The first step before the external data can be integrated is according to the best performing CRM analysis model (Engel et al, 2020), looking at the KPIs of the organization as starting point of the analysis. These key performance indicators should be determined based on what the management deems most important for the performance of the company. The key performance indicators will be

used to determine the current state of the company and will after the analysis and the integration of the solutions be used to evaluate the effect on the company. KPIs can be aspects of the organization like total profit or reputation score based on a survey. Once the KPIs have been determined a current state analysis of the organization can be performed. Here the company will score points on the KPIs and this will result in an overview of how the company is performing. These results can be used as starting point to choose what part of the organization the focus should be on and what external effects would be interesting to use, to influence the KPIs that are not performing well.

The next step is preparing the current database for the analysis. This is highly important, because if the internal data is not presented properly, then the inclusion of external will have little to no effect. This is why the database should be prepared using the techniques from chapter 2. The first aspect of the database that should be optimized is the adaptability. Use an analytic tool that is able to process different kinds of documents to improve the adaptability of your system. The adaptability is not only important for the insertion of the internal data, but external data also comes in different shapes and sizes so for the inclusion for both data types the adaptability of the chosen or made tool should be high.

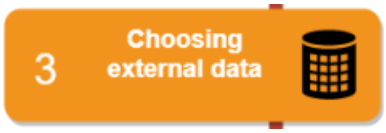
Another important aspect of the tool is its repeatability. Not only should the same results be presented if the analysis is done multiple times, but it should also produce the same results if the analysis is done by different people. Make sure the tool is easy to understand and learn. Provide training to the people that work with the tool to increase the repeatability. Not only should the tool be easy to understand to use but the results that are produced by the tool should also be easy to understand. Tools that produce easy to understand results have a high repeatability making it easier to translate the results into procedures that can be implemented within the organization to improve the KPIs.

One of the most important aspects of the database that came forth from both the literature review and the interview is data cleaning. The results of the analysis are less reliable if the data contains errors. There are a few methods that can be used for data cleaning which came forth from the interviews. Standardizing the input fields is the first step towards better data cleaning. The information that employees have to fill in during their work should be restricted to not too many options. The best way is to create dropdown field to ensure that no spelling errors are made or abbreviation are used. This is however not always possible if there are too many possibilities for a field. That is why the employees that work with the system have to be trained to learn what type of data they have to insert and the importance of writing down the data in the correct way. Be showing the employees that work with the system the added value when the data is inserted in the correct way, the willingness to make fewer mistakes increases. Even better is to involve the employees during the decisions and implementation of the analytic tool (Engel et al, 2020). By making use of a pilot for the chosen analytic tool not only the technical aspects of the tool can be evaluated but also the reaction of the employees to the new system. If the pilot works for a small group of employees and they enjoy the system then colleagues will most likely also adopt the system according to the ‘coaltion-of-the-willing’ theory which states that if management is able to convince a group of employees that the new system is a good addition, they will become early adopters and other colleges will follow once they see the early adopters using the new tool. After all these preparations errors in the data might still occur, that is why people within the organization should be tasked on

checking the data on errors. These errors can be found by making use of filters. If abnormalities are found they should be checked if it is an error, if so the data should be removed or transformed. Important for data removal and transformation is keeping track of who is adjusting the data and when, this way mistakes can be lead back to their origin. The last way to keep the data clean is not saving too much data. During the first step choices have been made on what the managements find important to analyze using the KPIs. Focus on saving the data that influences these KPIs and not too much else. If employees have to insert more data, the chance of errors becomes higher. Save the data also in as view tables as possible to keep the data structured and clean.

The last technical aspect of the internal tool of the organization that should be inspected is the degree of interpretability and robustness of the tool. Likes discussed earlier in the report interpretability means the degree to with users can understand the decisions that have been made by the system. The higher this degree of interpretability, the easier it is for the user to understand the prediction that has been made by the system. The robustness of the system, means how effective the algorithm of the system is in handling datasets. The more effective the system is at generating prediction models using different types of data the higher the robustness of the approach is. Both should be high in order to get the most out of the internal/external analysis.

Both the Business side and the technical side should be prepared according to the steps above to move on to the next step. Make sure to not focus too much on only one of the two sides, this has been the case in multiple models in the past and has led to less successful improvements of the organization though CRM-analysis. In case the business and technical side are up to standard then the next step is choosing what kind of data will be used for the external data.



Choosing the external data should be done through a series of interviews with the organization that wants to implement the external data. Import for selecting the external data is keeping in mind the connection between your own data and the external data in de database. For example, during the Van Eijck case the weather data was connected with the internal data using the latitude and longitude of existing incidents. Keep this in mind while selecting the data, that it fits with the current data. Extra data can be added to the internal data to make the external data fit but this increases the time and costs of adding the external data to your data base.

If you have no idea what type of external you want to add to the database, then there are a few public internet libraries to explore to find interesting data. These libraries are: Government sources like the CBS in the Netherlands, corporate filings like exchange commissions (SEC), Trade/business/professional associations, media, universities and foundations. These sources are a great starting point to define what kind of external data you want to add to your database. If the type of external data has been found then there are possibly multiple sources from which the external data can be extracted, choosing which source to use is the next step.

Next up is choosing which source should be used for implementing the external data. The literature has learned us that the three most important aspects of the external data are the flexibility, quality

and costs. Flexibility stands for the ease of use of the data. There are multiple methods for extracting external data like web services, flat file data transfer and web-scraping. The organization should decide how much time they want to lose on getting the data in the database. The logical answer would be as least time as possible, but lowering the flexibility often influences the costs and quality. For example, the Web-scraping method has a low flexibility but can be performed by your own employees who might perform at a lower cost than the costs of the web service integration and maintenance. The next part of the decision step is looking at the quality of the external data. The focus points of this qualitative analysis should include the following four points:

1. Relevance, which is the degree to which the data meets user needs both in coverage, content and detail.

2. Accuracy, which is the closeness between an estimated result and the (unknown) true value. Looking at how the data has been measured to see how accurate this method is

3. Timeliness, which is the degree to which data produced are up to date

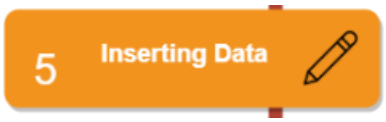
4. Comparability, which is the degree to which data can be compared over time, spatial domains and sub-population.

By looking at these four points for the different data sources, the quality of the data specifically for your organization will become clear. Like the flexibility it is also the case that a higher quality comes at the cost of the flexibility and actual costs.

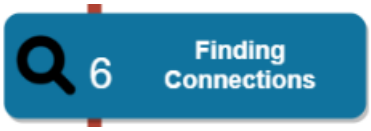
The last focus points of the source of the external data is the costs of the external data. These costs do include buying the data, training employees to use the data, maintenance costs of the data and potential costs of a larger database hosting. The best questions to ask yourself when examining the costs are:

- What is the business impact of incorrect data?
- What is the cost of maintaining the data feed?
- How large are the datasets?
- How often does the data change?
- How often does the data schema change?
- How complex is the data?
- How complex and varied are the consumption scenarios?
- What is the quality of the data (how many errors expected, how often, magnitude of impact)?
- How critical is the data to decision making?
- What are the auditing and traceability requirements?
- Are there any regulatory concerns?
- Are there any privacy or confidentiality concerns?

The organization should, based on the flexibility, quality and costs determine what the best source is for the external data. This decision can be made based on the KPIs which have been stated at the beginning of the approach, looking at the available money for the external source inclusion and how much they expect to get from it. For the flexibility and quality it is important to look at the knowledge which is in-house to determine how flexible and of what quality the data should be.

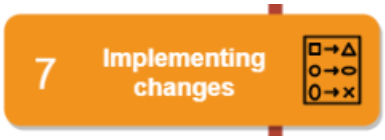


If the source has been chosen then the next step is to insert the external data into the database. This can be done in a number of ways. This can be on the way in using a vendor, integrating it at the core of the model, creating data marts for the external data, using a BI tool to combine the internal and external data or compare to different data bases. Based on the interviews about combining internal and external data, BI tools are become cheaper and are the easiest way to work with. However, there is not one best way to implement the data. The different stages of data inclusion all have their own benefits discussed in section 4.3. For example technical companies might prefer to create the analysing tools themselves and keep the user friendliness lower to keep the costs low.



If the external data is added to the database it is time to find the connections between the internal and the external data. The first step is to interview the company where the external data will be implemented to find what key events have happened in their company and what the key elements are of their system. During this interview it is also important to find out what the assumptions are of the organization towards the analysis. Make an overview of the key elements from this interview to create a clear structure of the most important parts of the database. The next step is creating visualizations based on the assumptions that the organization has of the influence of the external data on their company. These visualizations could be column charts and line charts. Using filters and slicers the visualizations can be adjusted to view the specific influence of a certain state of the external data. For example, during the Van Eijck case comparisons were made to incidents that happen in general and incidents that happen below a certain temperature. The same can be applied to other cases like income only below a certain amount of money. These filters must be adjusted based on the interviews with the client. If the visualisation shows anomalies in the data with the filters, it can be investigated further. The significance of the relationship should be tested for the next step. BI tools provide key influencers visualizations that should be used, these visualizations perform a test to see if the p-value is below 5% and can determine the exact correlation between the investigated object and its influencers.

If a BI tool has not been chosen for the insertion of the external database, then there are other type of analytical software like SPSS which can be used to confirm the relationship between different items. Once the key elements have been inspected with the use of the external data, you must have an interview with the company under research again to show the finding and how these have been found. It will often be the case that the company understudy will find a few errors in the analysis because certain assumptions have been made wrongfully or elements have been interpreted wrong. These mistakes should be fixed and the analysis should be reviewed again with the company under study.



The next step is to perform the actual changes within the company based on the analysis. Once the analysis has been successfully completed with the right interpretation of the data, changes can be made to the actual company. When implementing changes be sure to inform the employees based on the data why changes are being made and if needed offer training to let the employees learn to work with the changes.

During the next quarter after the changes have been implemented within the organization, the evaluation can start. The organization can analyze whether the changes to the company have positively affected the KPIs. Perform a new current state analysis and compare this with the one that has been made before the changes within the organization. In case the changes have not affected the company it is possible to go back to step six, to find other connections which can be improved to positively influence the company.

Chapter 6 Evaluation approach

The approach of chapter 5 which has been made using the literature review, the interviews with expert and empirical data from the case study with Van Eijck has to be evaluated to see if experts agree with the practices that are proposed in the approach and whether they think it will be useable for them when doing their jobs. This evaluation is performed by using a two-step approach. The first step is a questionnaire-based evaluation (Spil, 2005). The second step includes an in-depth group interview with the participants in the questionnaire-based evaluation. This step was done in order to share insights and perceptions of the approach, which were not covered in the questionnaire. A separate interview has also been held with a professor in the field of industrial engineering, to get an academic view on the current approach. The professor was unable to fill in the UTAUT questionnaire but was able to share his opinion on the approach for external data integration in its current form in figure 40.

6.1 Questionnaire

The questionnaire that is used is the UTAUT questionnaire, (Venkatesh and Davis, 2000). This questionnaire focusses on the performance expectancy of the approach, the effort expectancy, attitude towards using the approach, facilitating conditions, self-efficacy and anxiety towards the approach. The questions in the UTAUT questionnaire are built upon multiple theories from other literature like (Thompson, Higgins & Howell, 1991) and (Moore & Benbassat ,1991). The questionnaire has been filled in by five people working with CRM systems, which have also been interviewed as a group based on the questionnaire results. The results of the different elements of the UTAUT questionnaire are discussed in the sections below. The full filled-in questionnaires can be found in Appendix B.

6.1.1 Performance expectancy

The performance expectancy looks at the perceived usefulness, relative advantage and outcome expectations of the approach. All five participants think the approach would be useful for their jobs, that they could complete their jobs quicker if they made use of the approach and that it would increase their productivity. This means that the interviewees are positive about the possible performance of the approach.

6.1.2 Effort expectancy

The effort expectancy takes a look at the perceived ease of use of the approach and actual ease of use. All five participants agree that the approach is clear and understandable, they also think that it would be easy to become skillful in using the approach and that the approach is not difficult to use. This means that also towards the effort expectancies all five participants are positive.

6.1.3 Attitude towards using technology

The attitude towards using the technology looks at how willing the participants are of using the approach based is measured by the attitude towards behavior, the attitude towards the use of the approach and the effect of the approach. From the questionnaire, it can be concluded that the interviewees think that using the approach is a good idea and that they would like to work with the approach. However, they are neutral on the statements that the approach makes work more interesting and fun. Follow-up questions can be asked in the group session to find out how the approach can become more interesting to work with for the interviewees.

6.1.4 Facilitating conditions

The facilitating conditions look at what the participants need besides the explanation of the approach to make the approach work this is measured by the perceived behavioral control. The

results show that all five participants think that the created approach is compatible with other approaches that they are currently using. They are however not sure if they have the necessary resources and knowledge to use the approach at the moment. This can be solved by providing training to teach the approach to the company that wants to use it. Providing a list of resources in the approach can solve the problem that the users are not sure whether or not they have the resources needed in their company to use the approach.

6.1.5 Self-efficacy

Self-efficacy looks at how much help someone would need to use the approach using theories from Compeau & Higgins, 1995. The previous section about the facilitating conditions has already shown that the interviewees expect to need help when using the approach, how much help is further explored in this category. The results that the interviewees are not sure if it would help if they would be offered extra time to learn the approach. They do however think that they would be capable of using the approach effectively if they would be able to call someone that can answer their questions regarding the approach. This means that for the approach to be effective in an organization, certain people should be given a more in-depth training than others making it possible for them to help their colleagues if needed.

6.1.6 Anxiety

The anxiety towards using the approach looks at how intimidating the participants think the approach is. This is also measured with theories from Compeau & Higgins, 1995. The results from the questions about anxiety show that the interviewees do not feel apprehensive about using the approach. They do also not fear that the approach will let them make mistakes that they cannot correct and are not intimidated by the approach. This means that the interviewees feel comfortable while using the approach and are not overwhelmed by it. This was the goal of the approach to be seen as a usable option for organizations when it comes to combing CRM data with external data sources.

6.2 Interview results group session

Based on the results from the questionnaire, a group interview has been held with the expert on the field on CRM that have filled in the questionnaire. This was a semi-structured interview (Fontana & Frey, 1994) with as goal to find more points to improve upon for the approach. The interview gave the participants of the UTAUT questionnaire the opportunity to explain the answered that they gave and discuss how the approach can be improved. The focus points that were addressed besides the input of the participants were the understandability of the approach, the evaluation step in the approach and the business analysis of step one.

The first point of attention from the experts was *the current order of the steps in the approach*. At the moment step two is the preparation of the technical side. The experts thought that it would be better to move this step up to a phase if the type of external data is already known. They proposed to place it after step three or even after step four. This would mean that the preparation of the technical side will be done once the external data source has been chosen. This makes it easier to create or change a technical system for a client since it can be built with the external data in mind and the source from which it is coming. This source is especially important because it makes it easier to determine how flexible the system of the client must be. Meaning what types of data can be imported into the system. This is a good recommendation that will be applied to ensure that the technical preparation can be performed in a better way.

The next point of discussion was point eight, *the evaluation*. Currently the approach is linear. Based on the conclusion drawn in the evaluation, loops should be integrated into the approach. These

loops could lead to the preparation of the technical side if during the evaluation it becomes clear that the analysis has failed due to problems with the technique such as processing power. It could loop back to the decision of which external data to use. This would be the case if the connections between the internal and external data cannot be made easily or the external data is not complete enough. In this case another external data source should be found which provides a better data set. The last-mentioned loop point was from the evaluation back to finding the connections. This can be the case if the found connections have not been analyzed well enough or they have been but the resulting changes in the company have not caused enough improvement, so other connection possibilities should be investigated.

Another discussion point was that *only the evaluation of point eight is not enough*. During the approach and especially the finding connections phase the found connections should be validated to ensure that going back to step six is not necessary. This validation should be performed on a regular basis to see if no wrong assumptions have been made which could result in wrong elements selected or wrong filters. Performing this regular validation also ensures that the client stays more connected to the project and is therefore more likely to be happy with the end results since the client has been able to give feedback during the entire project.

The fourth point from the experts was that from experience they could also say that most companies have already established their KPIs. Meaning that point one, the preparation of the business side does not have to be fully performed at every company. They did however mention that in case the company already has KPIs, that it is a good idea to validate these KPIs based on experience in previous cases and by performing interviews to *see if the KPIs are representative* of the views of the organization. There should also be an evaluation at step eight, to see if the KPIs are still relevant to the company. The full implementation of external data and performing the changes in the company can be a lengthy process. So it might be the case that KPIs have changed over time. By checking this, it can be made sure the changes into the organization are not made towards the wrong goals.

When we spoke about the lengthy process of introducing external data to a database and performing changes in a company based on these changes another comment of the experts was *to write at every step how long this step is likely to take*. The experts expect that the hardest step is to find what external data the client exactly needs from which source and what they want to find out. This is something that they would like to see back in a new iteration of the approach. Showing the time of every step also improves the approach on a sales technical basis. Clients get a better idea of how long it will take before the external data is integrated with their data and how long it takes to find the relevant connection. By knowing how much time it will take and what the impact will be of the step, the client can make a better assessment of whether following the approach is worth it for them.

Another way of improving the sales technical part of the approach is by not looking at what external data to use in step four together with the client, but searching for external data that is usable for the client before reaching out to them. If a consultant can already tell the client what external data would be relevant to them which are available at sources the consultant knows, the client is more likely to agree to use the approach together with the consultant. For existing clients more knowledge is known about their database so they are a good starting point, by determining what external data could improve their business and proposing the approach to them. After getting more experience in using the approach at current clients, the consultant can move on to new clients. Possible new clients can be investigated to find what external data will fit into their organization. If a

business case is made for these clients based on possible external data, the chance is higher that they accept the external data project.

When discussing the use of the approach the experts agreed that learning how to use the approach efficiently would be a matter of time. A small training course would be enough to learn the basics of the approach and after this, they were certain that they would become better at using the approach over time. Having one person within the organization who is an expert at using the approach would increase the likelihood of using the approach successfully even more according to the experts.

The last comment that the experts had was that the approach could in the future be expanded or changed to be not specifically used for analysis with external data, but that it with some changes also can be used as an approach for data analysis in general. This means that organizations could use a smaller version of the approach that only focuses on the preparation of the analysis and the actual analysis to improve their company in a structured way even without the addition of external data.

6.3 Interview results professor

A professor in the field of industrial engineering has been interviewed in an unstructured interview. The professor sent the current approach before the meeting and during the meeting he shared his opinion on the approach and how it can be improved.

The first point was that the current approach could use a *bigger focus on data warehousing and filtering*. Currently the focus in the report is already for a big part of the analysis, but it is thought that a more elaborate explanation is needed to make sure the user of the approach needs less help when using the approach. The following slide has been sent from a lecture has been sent that could be added to the approach in future iterations.

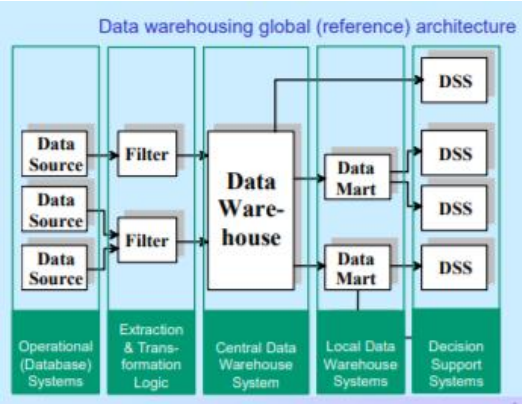


Figure 41: Data Warehousing information

The professor thought that the inclusion of the five steps from the data warehousing architecture: Operational, Extraction/transformation, central warehouse system, local warehouse systems and decisions could be elaborate more to give the user of the approach a better understanding of how to combine internal with external data.

The second point of feedback was that the business preparation could also be made more elaborate than its current form. This was also discussed in the group interview meeting in the previous section 6.2. It should be explained that most organizations have already established KPIs, so if an organization does already have established KPIs these should be investigated rather than creation always new KPIs.

The third feedback point was on *the current place of step three*: Choosing the external data. Just like the group interview, the professor also thought that choosing the external data should be at the beginning of the approach to improve the preparation of the technical and business side of the project. This could be done by creating a business case before reaching out to a client, this was also found in the group interview, so is certainly an important point for future iterations of the approach in further research. It was mentioned that good preparation is half the work.

The fourth point of attention was that it should be *clearer in the approach what sources were used* for the creation. Without reading the entire master thesis, it seems according to the professor that too much of the approach was created by empirical data from the case study. The wheel should not be reinvented. Step 1,2,3,4 and 6 are mostly created with the help of existing literature, so this should be made clearer in future iteration. Without making clear that this approach is not created out of thin air but with the help of current literature, researchers might avoid using the approach.

The last feedback point was that this approach could mostly be seen as a *general instruction* on how the organization can choose, integrate and analyse external data combined with internal data and that there are more specific existing approaches that go in depth on the separate points of the approach created during this research. For example, the user of the approach might not have any trouble following the steps until step four: Choosing the source, then he should be instructed to separate existing approaches that focus only on the choosing of the data source to help him. The power of the approach created in this research would increase if separate approaches are suggested within this approach in case the user gets stuck. Possible approaches that can be combined with the approach created for this research will be discussed in contribution part of this research.

Chapter 7 Conclusion

The objective of this research was to answer the main research question: **What is an appropriate approach that organization can use to perform a tool based analysis to improve processes using external data integration based on current literature, experts and a case study?** This has been done by the creation of an external data integration approach in figure 40. This approach has been created based on the answers from the eight sub-research questions. Answers were acquired by performing a literature research, performing interviews and empirical data from a case study at Van Eijck. The answers from all the sub-research questions and the answer to the main research question will all be discussed in the discussion. Recommendations to Van Eijck and Lumen BS will be given as well as contributions to current theories, ideas for future research and the limitations of this research.

7.1 Discussion

RQ1. What is the state of art of process optimization using system-generated data?

To get the most out of internal data in combination with external data, it was first looked at what the current state of art is for process optimization using system-generated data using a systematic literature research. This literature research looked at multiple models about system-generated data and a ranking of the best papers was made in section 2.2.3. This ranking was made based on the most important aspects of data analysis found during the literature research. It was found that most models for process optimization using system-generated data focus mostly on either the technical side or the business side of the optimization and only a few focused on both these aspects. A reoccurring elements in most models was the inclusion of data cleaning, adaptability and repeatability. These technical elements seem to be essential for process optimization using system-generated data according to current literature.

RQ2. What are the elements which should be focussed on when improving process optimization using system-generated data?

To find the elements which should be used in the approach for external data integration created in this research the found aspects of the models from RQ1 were investigated to find out what elements improve process optimization using system-generated data and what should not be included in the final approach for external data integration. The most important aspects a model should include are on the business side: KPIs and a Business analysis. On the technical side the model should have: adaptability, data cleaning and repeatability. Adaptability stands for how many different data types can be included in the analysis of the system. Data cleaning looks at the functions that are available for transforming the data in such a way that there are no longer mistakes in the data and repeatability stands for the degree to which the same results are generated if the analysis is done multiple times and not always by the same person. The more aspects the model includes the stronger it is towards data optimization with as goal optimizing processes within an organization. The paper that came up on top of the ranking was Engel et al(2020). This paper included all the important aspects of data analysis both from a technical viewpoint and a business viewpoint.

The weakness of most models was that they only focused on either the technical or the business side of the process optimization as can be seen in table 5 in section 2.2.3. From the eight models that were investigated during the research only two papers looked at both the technical and the business side of the analysis, while the combination of these two gives a model real strength and usability for a company. From this literature was learned that the approach generated by this research should focus on both the business and technical side of an organization while preparing them for the

analysis with external data. During this preparation the strong points of section 2.3.1 should be included to make sure the chance increases that a successful process optimization is performed.

RQ3. What is the current best-practice in industry to find connections between internal and external data?

This question has been answered by performing interviews with experts in the field of data analysis, the results of these interviews can be found in section 3.4. From these interviews was concluded that to find connections between internal and external data, the quality of the internal data must be as high as possible. This can be done by making use of input fields and standardizing the data. All the people that insert data into the database should also use the same standards to avoid confusion. It is also important that not only the technical system is made in such a way that the data is kept clean, but that also the employees try their utmost best to keep the data clean. This can be done by offering pieces of training to the employees to teach them about the value the data has for the organization. Once they see the reason why they have to insert the data, the chance becomes smaller that they will make use of abbreviations according to the interviews. Another way of reducing the human errors in the data is making sure that it is saved who inserts and transforms which data. This way it can be lead back to who created the errors in the data.

The ways that the internal and external data can be analyzed together according to the experts is by using special databases with time/space analysis if the researcher is familiar with this kind of technology and wants to save money. The more expensive method but visually more pleasant one is using a BI tool. These are still not really expensive but offer the user a more intuitive system for combining internal and external data. The functions of the BI tool can be extended with other tools like Azure. If these tools are used and the data is kept clean using the methods in section 3.4, data connection between the internal of an organization and external data are more likely to be found.

RQ4. What goals do the stakeholders of the Van Eijck case of optimizing processes using weather data have, which is used for the development and evaluation of the external data integration approach?

The approach for external data integration has been made based on empirical data from a case study. This case study was performed with Van Eijck, a salvage company that helps cars when they have had an incident on the road. Van Eijcks CRM database has been enriched with external data for the case study. To get the most out of this case study, a stakeholder analysis was performed to find out what the goals of each stakeholder were. This was important to find out because it made performing the analysis easier, since the end goal of the analysis became clear.

The four stakeholders for this case study were Lumen BS, Van Eijck, Rijkswaterstaat and weather data providers. All these stakeholders have been analyzed using a stakeholder analysis matrix in section 3.2 table 8. With the information gained from this analysis the stakeholders were also placed on the power-interest matrix in figure 5. This figure has taught us that Lumen BS and Van Eijck are the most important stakeholders. The goal of Lumen BS was to find a way to implement external data at their clients and potential clients, this will be done by using the approach created for the main research question. The goal of Van Eijck was to reduce the time it took to arrive at an incident, ways to reduce this time are discussed in section 4.6 and 7.3.1.

RQ5. How can we integrate external data and analyse it based on the case study in such a way that it takes less time and effort than is currently needed for the integration and analysis?

This research question has been introduced to find out what steps should be included in the final approach for data analysis with external data to improve processes within an organization. This research question has been answered in chapter 4. The tool that was used in this chapter is Power BI combined with Microsoft dynamics 365. It was chosen to use this tool because Lumen BS had already provided these tools to Van Eijck before the research and they were familiar with it.

The practice that can be used to reduce the time and effort of data analysis and integration is using the critical realism approach to find connections in the data and integrate the data with the correct connections in a structured way. First a description of the events was given in section 4.1 followed by the key elements which were important for the analysis in section 4.2. After these two steps another literature review was held in section 4.3 for the theoretical re-description. This section looked at what aspects to focus on when adding external data to your database, because no relevant literature was found specific for the salvage sector that could be used in the analysis. The entry point of external data is important for the use of the tool, because the re-description shows that the external data in a data analysis can be inserted at multiple points, each point changes the way the tool is used. At this point in the research, the external data was already added to the database, but the lessons learned from the theoretical re-description have been used for the creation of the approach that answers the main research question.

The use of the tool is most visible in section 4.4, which is the retroduction part of the critical realism approach. This section shows the multiple visualizations that have been built in Power BI to find relevant connections in the data of Van Eijck. The way the Power BI tool was used in this step was first creating graphs based on the interviews with Van Eijck which gave a general idea of what datatypes influenced the number of incidents and the arrival time at incidents the most. After this step a significance test was performed in Power BI using a key influencer visualization, which presented relevant data connections based on a significance test that looked at the size of the dataset that was being investigated and if the connection was the internal and external data was not random or caused by other factors.

By using the critical realism approach and looking at significant relationships in the data while using the tool, multiple relevant relationships between the internal and external data were found with less effort and time needed than it would have taken without this structured method. This information is used for the analysis part of the approach that has been created.

RQ6. How to design an approach that fits the goals of the main research question based on current literature, experts and the case study?

To answer the main research question an approach for external data integration should be made that companies can use for process optimization. The first iteration of this approach has been created using the answers of the previous research questions. The methods used for designing this approach are a systematic literature review, interviews with experts and empirical data from a case study. With the help of these methods the approach in chapter 5 figure 40 has been created. The answers from the previous research questions resulted in the approach existing out of eight steps. These are: The preparation of the business side, the preparation of the technical side, choosing the external data, choosing the source, inserting the data, finding the connections, implementing changes and performing an evaluation. The first four steps from the approach have been designed using the findings from the literature research in chapter 2 and the interviews from chapter 3. The

last four steps have been created using the empirical data from the case study of chapter 4 and current literature. This combination of sources led to the first iteration of the external data integration approach. A more in-depth explanation of each step is given in chapter 5.

RQ7. How can the usability of the proposed approach be improved based on an evaluation of experts?

The approach which has been created using the literature, the interviews and the empirical data from the case study has been shown to experts. These experts have given their opinion on the approach using the UTAUT questionnaire, these can be found in appendix B. The UTAUT questionnaire was chosen because the approach for external data integration is a new artifact that organizations can use when implementing and analyzing external data and using the UTAUT questionnaire it was investigated whether organizations would be willing to use this new artifact. Based on this questionnaire a group interview has been held with all the participants that filled in the questionnaire, the results of this interview can be found in section 6.2. A professor from a university has also been interviewed to get an academic perspective on the current approach created for RQ6.

The interviewees were positive about the current structure of the approach but thought it would be better if the technical step was moved to another position in the approach after the external data type is known. They also would prefer it if there would be made loops in the approach based on the outcome of the last step, the evaluation. Another proposal by the experts was to increase the usability by improving the approach sales technical. This can be done by not waiting on step three of the approach to discuss with the client what possible external data would be useful for their business, but by creating a business case about the best external data for their organization before reaching out to the potential client. This will increase the chance that the client is willing to work together and follow the approach.

Another method to increase usability is to perform training courses at the organization that wants to make use of the approach. The experts thought that they would need a short training before they could work with it and after this they would become better at using the approach by experience.

The professor has added to this feedback of the experts from organization that the current approach will improve academically if the current explanations for each step are extended with more references on why the suggestions are given. The approach should also provide links to specific existing approaches if a user of the approach gets stuck during one of the steps. For example if they are stuck at inserting the data, it would be helpful if the approach from this research would suggest an approach specifically for the insertion of data from current literature. The full list of improvement points for future iteration can be found in section 6.2 and 6.3.

RQ8. How can the proposed approach help organizations in general improve their processes with external data compared to existing external data integration approaches?

The current approach for external data integration in chapter 5 has been created using a case study from a service provider company and made use of literature from CRM-system analysis for process optimization. Based on the interviews in chapter 6 it was found that the approach can also be used at other sectors and is not restricted to CRM-system data as internal data set. When working together with organizations from other sectors it is important to research the sector that this organization is from and create a business case to improve the chance that the approach for external data integration will fit at the new organization and that they will make use of it according to the results from section 6.2 and 6.3. The most important step for this business case is finding out what

kind of external data is most fitting for the company that you are trying to help with the external data integration approach.

The experts also determined that the approach is not restricted to CRM data but can also be used at data from other sources. Lumen BS also helps with ERP systems and the experts thought that the approach could also be used for the data generated by these EPR systems. As long as a connection can be created between the internal data and the external data and a business case is created as first step of the approach, the approach can be used more generally.

Based on these eight research questions the main research question: "**What is an appropriate approach that organization can use to perform a tool based analysis to improve processes using external data integration based on current literature, experts and a case study?"** has been answered by the creation of the approach in chapter 5 together with the feedback points from chapter 6. This approach has been created using the literature research from chapter 2, the interviews of chapter 3 and the empirical data from the case study which has been performed in chapter 4. Figure 40 shows the approach which organizations can use to prepare their organization for external data, find external data and sources which fit their company, find connections between the internal and the external data and make the changes in the company. The eight steps of this approach are fully explained in chapter 5. In short, an organization should first focus on preparing the business and technical side of their company by following the strong points of system analysis models from RQ1 & RQ2 and the interview findings from RQ3. After this the external data, its source and the implementation place should be chosen following the practices of section 4.3 and the answer of RQ5. This will result in the database being enriched with external data. The next step is to perform the analysis following the steps of chapter 4 in which the significance test for the relationships is of utmost importance for a successful analysis. Changes in the organization should be made based on the findings in the analysis and when the changes have been completed, the whole process should be evaluated following the explanation in the last step of the external data integration approach.

Based on the questionnaire and the group interview with the data analysis experts there are a some changes that should be made to this approach in following iterations to make it even more useful. These are discussed in section 6.2 and the answer on RQ6 in this section. The most important ones are that the choice of the external data should be done as the first step in the form of a business case to convince clients to use the approach and evaluations should be held at every step to see if everything is still in line with the KPIs, if this is not the case then the user of the approach should loop back.

The created approach that answers the main research question is an approach that organization can use together with existing approaches which focus solely on one or two steps of this approach. This approach from figure 40 is a general guide for organizations that they can use when they start from scratch and want to use external data to improve their organization, but shows mostly the basic principles of every step. A more in depth explanation can be found by reading this master thesis and by making use of existing approaches focused on for example only the insertion of external data in a database. The quality of the approach would grow if every step would include suggested existing approaches in case they would get stuck. Possible other approaches that can be used in combination with the approach from the main research question will be discussed in section 7.2.

7.2 Contribution

This research has created value for Lumen BS and Van Eijck by providing Lumen BS with an approach which they can use at other clients when performing analysis with external data integration and

provided Van Eijck with recommendations in section 4.6 and a dashboard which they can use to analyze their data even further. The results from this thesis are also an addition to current theory, namely the approach that has been created for external data integration. This section will look at how the proposed approach contributes to other approaches from existing theory.

When performing the research for this master thesis, it was found that there are no existing analysis approaches online which focus on the integration of external data from beginning to the end. Current theory exists mostly out of either approaches for data analysis or approaches for the integration of data. The combination of the two is a new addition to current theory. An example of an approach for data integration is the essential steps in the data integration process by Athena IT solutions. This approach focusses on finding the right business requirements by looking at the costs and quality of the data like step three from the external data integration approach in figure 40. The approach by Athena solutions also looks at the preparation of the system and business, data franchising and data management. This approach goes a little bit more in-depth at preparing the technical side of the data integration. Another example of an approach for data integration is provided by data integration info(Ashraf, 2020). Here the focus is again more on the technical side. This approach explains in six steps how to extract data. This approach is also a more in-depth version of step five from the external data integration approach. The last found approach that focuses mostly on the data integration is by ETL solutions. This approach is covers the business side of integrating data in a project, explains how external data should be chosen and how it should be imported into the data base of the organization. These three approaches give more theory on how the user should integrate external data into a database than the approach created in this research, but the approach from this research provides the user with more information of not only integrating data into the database, but also making use of this data to improve the organization.

The same applies to the analysis part of the approach of this research compared to existing analysis approaches. An example of an existing data analysis approach (Oner, 2017), shows in six steps how organizations can transform the data that they own and analyze it, this approach focusses on the technical aspects for data analysis while other approaches like (Dillard, 2017) focusses more on the business questions which are important for data analysis. Like the integration theories explained in the section above, the two theoretical approaches for data analysis are also more elaborate than the explanation of the analysis in step six of the external data integration approach from this research. But here again it is the case that the approach from this research provides a more complete picture of not only the analysis but also helps with making the changes in the organization based on the analysis and the evaluations that need to take place after the analysis.

The approaches that come the closes to the approach created for this master thesis are the big data approach of the Dutch chamber of commerce and the CRISP-DM approach for data use. The approach of the Dutch chamber of commerce focuses on how to find external data and analyze it together with internal data, but does not look at the evaluation and implementation part of the research. The analysis part of this approach is also very limited. The CRISP-DM approach is the approach that was found that is the most similar to the approach created during this research. This approach follows a cycle that looks similar to the steps from figure 40, but CRISP-DM is not made with external data taken into account and does therefore not explain to its user what the focus points are when selecting and inserting external data.

Since there have been no approaches found in current theory which cover the entire integration part of external data, how to analyze it and how to perform the changes, this approach can be seen as new piece of theory that organization can use when they want to make use of external data and want a complete picture of the whole integration and analysis. The steps in the approach are

explained in a general sense but like showed by the previous example approaches of data integration and analysis are well complemented by current theoretical approach. Researchers can use the approach from this research as starting point when they want to learn more about external data integration and analysis and when they need a more in-depth explanation of one of the steps in the approach, they can look at one of the theoretical approaches described in the previous two sections. This was also advised during the interview with a professor on industrial engineering when he was shown the current external data approach from figure 40. This way the findings from this research fit right into the current set of knowledge on external data and give more value to other theories by showing researchers and organizations how they can combine theoretical approaches on the different parts of external data integration with as goal process optimization following the steps from the approach of this research.

Besides showing researchers how they can combine older approaches, this research also adds current theories from experts in the field of data analysis. Other researchers can use the knowledge that is documented from the experts on subjects like data cleaning to use in other research if they want to use up-to-date information about data analysis.

This research also provides a documented example of how data analysis can be performed in the form of the Van Eijck data analysis case. The empirical data from this case in chapter four can be used by other researchers if they are in need of an example of a data analysis case before doing a data analysis themselves.

The last addition to theory from this research is that this research also provides a more sales technical view on data integration and analysis because of the information that has been gathered during the interviews with experts in the field of data analysis in section 6.2. Current theory focuses mostly on how the integration and analysis can be performed as efficiently as possible but do not look at the sales technical aspect. The data from the interviews in section 6.2 does explain this and therefore provides information on a less explored view on data integration and analysis, the sales technical view.

To summarize the added value of the results from this master thesis to theory. The approach created to answer the main research question is an addition to the field of data integration and the field of data analysis. It can be used as a general guide when wanting to optimize processes within an organization and can be used in combination with other approaches like the ones from Oner (2017) and Ashraf (2020). Creating a better link between existing approaches which focus on either data integration or data analysis.

7.3 Recommendation

Following the discussion of the research questions and the contribution of this research to theory, recommendations have been made for the two main partners of this master thesis, Van Eijck and Lumen BS. These recommendations will answer the goals of these two organizations which have been found in section 3.1.

7.3.1 Recommendations for Van Eijck

The goal of Van Eijck is to reduce the time it takes for their employees to arrive at incidents. If this time is not reduced at certain rayons they are at risk of losing the rights to assist cars in these rayons. To avoid that, Van Eijck should start by following the suggestions of section 4.7. Here it is stated that a few rayons are significantly more affected by the weather than other rayons. The number of people and cars on standby in these rayons during bad weather should be relatively more than at other rayons. Van Eijck has said that they want at least incident data from three years before

acting upon it, but my recommendation is to already make a few smaller changes based on the suggestions in section 4.7. After three years it is expected that the database has information of more than 250.000 incidents, with enough data on each weather type to form a more meaningful conclusion. The visualizations created for Van Eijck shown in Figures 15, 17, 19, 21 and 23 can be used they inspect if each weather type has enough data. The results from investigating these 250.000 incidents can be used to create prediction models. These prediction models should look at the predicted weather condition and based on this recommend which rayons need more or fewer cars and employees, to ensure the arrival time at incidents is on time and that rayons are not filled with too many employees. Van Eijck could choose to add more external data like the quality of the roads to improve the possible prediction models even further and increase the chance of optimizing processes within the organization.

For this research a lot of visualizations have been built within the dashboard of Van Eijck which they can use to not only look at how the weather data influences their incidents and arrival time but also how the arrival times of every rayon are performing in general. The advice is that while making the reports they also look at these visualizations because they might explain anomalies in the data. It is also advised to keep building upon the weather data visualization, to optimize them even further. Following this advice increases the chance that the arrival time at incidents will reduce and in doing so optimize their business processes.

7.3.2 Recommendations for Lumen BS

The goal of Lumen BS is to find a systematic approach with which they can use to help their current clients or new ones when they want to enrich their database with external data. This approach has been created and can be found in chapter 5 figure 40. It is advised that Lumen BS uses this approach for future clients when they want to enrich their database with external data in combination with approaches discussed in section 7.2 if they need more information about a specific step of the approach from this research. This approach should also be used in combination with the advice provided by the experts of section 6.2 and 6.3. This means that to make the approach more attractive for clients they should make business cases on what external data would be useful for the client before reaching out to the client. The approach is usable at all clients for which Lumen BS can come up with a business case on how external data would be useful for the client, it is not restricted to organizations within the service sector or organization that make use of CRM systems to gather their information. They should also perform multiple validations throughout the process of the external data integration approach to make sure the process is still going the right way according to the client and no wrong assumptions have been made.

To get the most out of the approach, training courses should be offered within the organization on how to use the approach in the best way and efficiently. After this training the employees should understand the basics of the approach and can get better at using it through experience. If they are any questions regarding the approach, they can get into contact for questions regarding certain steps in the approach or make use of the approaches suggested in the contribution to get more details on certain steps of the approach created during this research.

7.4 Limitations

This research was bound by several limitations which will be discussed in this section. A few limitations specific to the literature research have already been discussed in section 2.4.1.

The first limitation of this project is that the approach created by this research has been made based on a case study at a single company. The quality of the approach created to answer the main research question would improve if the experience did not only come from a single organization but

from a number of companies. The same applies to the validation of the approach. This has been done by experts that work at Lumen BS (Seddon & Scheepers, 2012). People that work at the same organization tend to have the same view on things. This is why it would be better if in the future, experts from outside of Lumen BS take a look at the approach to evaluate it.

The second limitation of the research was that for the analysis part of the research in chapter 4, the focus was on the Power BI tool by Microsoft dynamics 365. This could have influenced the approach that was created, because step six, finding the connection of the approach is based on the experiences from the Power BI tool while other tools are also available to use for performing an analysis to find connections between data sets. During the creation of the approach it has been tried to make it applicable to other tools to improve the generalizability of the approach and while this has succeeded , it could be the case that because the empirical data from the case study was from power BI, that some small analysis points from the approach are harder to accomplish with other tools

The third limitation is the number of interviews held with the stakeholders in chapter 3. Two experts have been interviewed to give their opinion on the integration of internal and external data. By interviewing more people, the questions could be answered from more angles and more findings could have been made. Also due to having no contact person at the weather data providers, this stakeholder is the only stakeholder that has not been interviewed during the research.

The fourth limitation was that because of the coronavirus, it was not possible to be physically at the office of Lumen BS and Van Eijck throughout most of the research. Online meetings reduced intractability and made it harder to find out the real needs of Van Eijck. Process within this research could have been performed faster if meetings in person were possible.

7.5 Future work

This research can be extended in multiple ways, these will be discussed in this section.

First, I consider it important to test the approach at organizations besides Van Eijck, to see if it works as efficiently as expected. Based on this implementation of the approach, changes could be suggested to improve the quality of the approach. The organizations at which the approach is tested can be from different fields and should not be restricted to only the service sector like Van Eijck. By testing the approach at all kinds of sectors it could also be found whether the approach works for every sector or is restricted to a number of sectors.

The second line for future research is extending the functionality of the approach to not only be applicable to CRM-system data combined with external data, but that the approach can also be used at all kinds of data from other systems than a CRM-system and that external data could also be just other internal data. Experts from Lumen BS already thought that the approach has the possibility to be extended be also usable on these cases, but further literature research and test studies should be performed to find out the real extendibility of the approach made by this research.

The third future research possibility is creating a new iteration of the approach using the comments of the experts in section 6.2. At the moments the comments have been fully written out and compared with the current state of the approach, but future researchers could change the model depicted in figure 40 to include the comments of the experts on data-analysis.

The fourth future research possibility is that Van Eijck has made it known that they want to go further with the current visualizations that were provided to them in Power BI. These visualization focus however on weather information, but there are more types of external data that could be

usable for Van Eijck, like the quality of the road or the side of the road on which the incident has happened. Other researchers can work together with Van Eijck to implement these external data types into their CRM-system database and provide them with even more possibilities to explain anomalies and reduce the time it takes to incidents.

Lastly, the current approach can be improved by adding more clearly which approaches from section 7.2 belong by which step. It could be improved even further if more research will be done in the field of approaches for data analysis and data integration. This information can be used to improve the approach from this research directly or it can be used to provide the user of the approach with more links in case a step of the approach is not clear enough to use.

References

[1] Alshawi, S., Missi, F., & Irani, Z. (2011). Organisational, technical and data quality factors in CRM adoption — SMEs perspective. *Industrial Marketing Management*, *40*(3), 376–383. https://doi.org/10.1016/j.indmarman.2010.08.006

[2] Altalhi, A. H., AL-Malaise AL-Ghamdi, A., Ullah, Z., & Saleem, F. (2016). Developing a framework and algorithm for scalability to evaluate the performance and throughput of CRM systems. *Intelligent Automation & Soft Computing*, *23*(1), 149–152. https://doi.org/10.1080/10798587.2016.1184830

[3] Ashraf, S. (2020, 12 augustus). *Best Ways for Successful XML Data Integration*. Data Integration Blog. https://dataintegrationinfo.com/xml-data-integration/

[4] Bahari, T. F., & Elayidom, M. S. (2015). An Efficient CRM-Data Mining Framework for the Prediction of Customer Behaviour. *Procedia Computer Science*, *46*, 725–731. https://doi.org/10.1016/j.procs.2015.02.136

[5] Chalmeta, R. (2006). Methodology for customer relationship management. *Journal of Systems and Software*, *79*(7), 1015–1024. https://doi.org/10.1016/j.jss.2005.10.018

[6] Chanial, C., Dziri, R., Galhardas, H., Leblay, J., Nguyen, M.-H. L., & Manolescu, I. (2018). Connectionlens. *Proceedings of the VLDB Endowment*, *11*(12), 2030–2033. https://doi.org/10.14778/3229863.3236252

[7] Chen, I. J., & Popovich, K. (2003). Understanding customer relationship management (CRM). *Business Process Management Journal*, *9*(5), 672–688. https://doi.org/10.1108/14637150310496758

[8] Chen, J.-S., Yen, H. J. R., Li, E. Y., & Ching, R. K. H. (2009). Measuring CRM effectiveness: Construct development, validation and application of a process-oriented model. *Total Quality Management & Business Excellence*, *20*(3), 283–299. https://doi.org/10.1080/14783360902719451

[9] Chiang, W.-Y. (2014a). Applying data mining with a new model on customer relationship management systems: a case of airline industry in Taiwan. *Transportation Letters*, *6*(2), 89–97. https://doi.org/10.1179/1942787514y.0000000014

[10] Chiang, W.-Y. (2014b). mining high value travelers using a new model designed for online crm systems: a case study in taiwan. *Aviation*, *18*(3), 157–160. https://doi.org/10.3846/16487788.2014.969879

[11] Dillard, J. (2017). *The Data Analysis Process: 5 Steps To Better Decision Making*. Bigskyassociates. https://www.bigskyassociates.com/blog/bid/372186/The-Data-Analysis-Process-5-Steps-To-Better-Decision-Making

[12] Eid, R. (2007a). Towards a Successful CRM Implementation in Banks: An Integrated Model. *The Service Industries Journal*, *27*(8), 1021–1039. https://doi.org/10.1080/02642060701673703

[13] Eid, R. (2007b). Towards a Successful CRM Implementation in Banks: An Integrated Model. *The Service Industries Journal*, *27*(8), 1021–1039. https://doi.org/10.1080/02642060701673703

[14] Engel, M., Schoonderwoerd, R, (2020). Data-analyse binnen de transport- en logistieksector: nog veel onbenut potentieel, Mazars.

[15] Fontana, A. and J. Frey. 1994. The Art of Science." Pp. 361-76 in The Handbook of Qualitative Research, edited by N. a. Y. L. Denzin. Thousand Oaks: Sage.

[16] Guerola-Navarro, V., Oltra-Badenes, R., Gil-Gomez, H., & Gil-Gomez, J. A. (2020). Research model for measuring the impact of customer relationship management (CRM) on performance indicators. *Economic Research-Ekonomska Istraživanja*, 1–23. https://doi.org/10.1080/1331677x.2020.1836992

[17] Guo, F., & Qin, H. (2017). Data Mining Techniques for Customer Relationship Management. *Journal of Physics: Conference Series*, *910*, 012021. https://doi.org/10.1088/1742-6596/910/1/012021

[18] Henneberg, S. C. (2006). An Exploratory Analysis of CRM Implementation Models. *Journal of Relationship Marketing*, *4*(3–4), 85–104. https://doi.org/10.1300/j366v04n03_06

[19] Hosseini, S. M. S., Maleki, A., & Gholamian, M. R. (2010). Cluster analysis using data mining approach to develop CRM methodology to assess the customer loyalty. *Expert Systems with Applications*, *37*(7), 5259–5264. https://doi.org/10.1016/j.eswa.2009.12.070

[20] Kambatla, K., Kollias, G., Kumar, V., & Grama, A. (2014). Trends in big data analytics. *Journal of Parallel and Distributed Computing*, *74*(7), 2561–2573. https://doi.org/10.1016/j.jpdc.2014.01.003

[21] Khashab, B., Gulliver, S., & Ayoubi, R. M. (2020). Scoping and aligning CRM strategy in higher education institutions: practical steps. *Journal of Strategic Marketing*, 1–25. https://doi.org/10.1080/0965254x.2020.1823458

[22] Khashab, B., Gulliver, S. R., & Ayoubi, R. M. (2018). A framework for customer relationship management strategy orientation support in higher education institutions. *Journal of Strategic Marketing*, *28*(3), 246–265. https://doi.org/10.1080/0965254x.2018.1522363

[23] Ku, E. C. S. (2010). The impact of customer relationship management through implementation of information systems. *Total Quality Management & Business Excellence*, *21*(11), 1085–1102. https://doi.org/10.1080/14783360903250514

[24] Li, Y., Huang, J., & Song, T. (2019). Examining business value of customer relationship management systems: IT usage and two-stage model perspectives. *Information & Management*, *56*(3), 392–402. https://doi.org/10.1016/j.im.2018.07.012

[25] Löfgren T, Gravem D, Haraldsen G. (2011), A glimpse into the businesses' use of internal and external data sources in decision-making processes, Blue-ETS

[26] Ngai, E. W. T., Xiu, L., & Chau, D. C. K. (2009). Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, *36*(2), 2592–2602. https://doi.org/10.1016/j.eswa.2008.02.021

[27] Oner, T. (2021, 25 maart). *Verregaande data analyse maken in 6 stappen*. Traffic Builders. https://www.traffic-builders.com/data-analyse-6-stappen/

[28] O'Reilly, K., & Paper, D. (2009). Customer Relationship Management (CRM): An Approach for Transforming the 'Myth' of CRM Success into Dual-Creation of Value. *Journal of Information Technology Case andApplicationResearch,11*3),1–8.https://doi.org/10.1080/15228053.2009.10856162

[29] Peelen, E., van Montfort, K., Beltman, R., & Klerkx, A. (2009). An empirical study into the foundations of CRM success. *Journal of Strategic Marketing*, *17*(6), 453–471. https://doi.org/10.1080/09652540903371695

[30] Peter B. Seddon, Rens Scheepers: Towards the improved treatment of generalization of knowledge claims in IS research: drawing general conclusions from samples. Eur. J. Inf. Syst. 21(1): 6-21 (2012)

[31] Petschulat, S. (2010). Other people's data. *Communications of the ACM*, *53*(1), 53–57. https://doi.org/10.1145/1629175.1629196

[32] Pieragostini, C., Mussati, M. C., & Aguirre, P. (2012a). On process optimization considering LCA methodology. *Journal of Environmental Management*, *96*(1), 43–54. https://doi.org/10.1016/j.jenvman.2011.10.014

[33] Pieragostini, C., Mussati, M. C., & Aguirre, P. (2012b). On process optimization considering LCA methodology. *Journal of Environmental Management*, *96*(1), 43–54. https://doi.org/10.1016/j.jenvman.2011.10.014

[34] Rajput, A., Zahid, M. M., & Najaf, R. (2018). Using CRM to Model Firm Performance in a Business-to-Business Market. *Journal of Relationship Marketing*, *17*(2), 118–151. https://doi.org/10.1080/15332667.2018.1440142

[35] Richard, J. E., Thirkell, P. C., & Huff, S. L. (2007). The strategic value of CRM: a technology adoption perspective. *Journal of Strategic Marketing*, *15*(5), 421–439. https://doi.org/10.1080/09652540701726793

[36] Saldana, J. (2012). The Coding Manual for Qualitative Researchers. SAGE Publications.

[37] Samaranayake, P. (2009). Business process integration, automation, and optimization in ERP. *Business Process Management Journal*, *15*(4), 504–526. https://doi.org/10.1108/14637150910975516

[38] Shukla, M. K., & Pattnaik, P. N. (2018). Managing Customer Relations in a Modern Business Environment: Towards an Ecosystem-Based Sustainable CRM Model. *Journal of Relationship Marketing*, *18*(1), 17–33. https://doi.org/10.1080/15332667.2018.1534057

[39] Song, M., Zhao, X., E, H., & Ou, Z. (2016). Statistic-based CRM approach via time series segmenting RFM on large scale data. *Proceedings of the 9th International Conference on Utility and Cloud Computing*, 282–191. https://doi.org/10.1145/2996890.3007873

[40] Spil, T. A. M. (2005). *E-health Systems Diffusion and Use* (illustrated edition). Igi Global.

[41] Stein, A., & Smith, M. (2009). CRM systems and organizational learning: An exploration of the relationship between CRM effectiveness and the customer information orientation of the firm in industrial markets. *Industrial Marketing Management*, *38*(2), 198–206. https://doi.org/10.1016/j.indmarman.2008.12.013

[42] Valmohammadi, C. (2017). Customer relationship management: Innovation and performance. *International Journal of Innovation Science*, *9*(4), 374–395. https://doi.org/10.1108/ijis-02-2017-0011

[43] Valos, M. J., & Bednall, D. H. B. (2010). The alignment of market research with business strategy and CRM. *Journal of Strategic Marketing*, *18*(3), 187–199. https://doi.org/10.1080/09652540903537022

[44] Venkatesh, V., and Davis, F. D.(2000) "A Theoretical Extension of the Technology Acceptance Model: Four Longitudinal Field Studies," Management Science (45:2), pp. 186-204.

[45] Volkoff, O., Strong, D. M., and Elmes, M. B. 2007. "Technological Embeddedness and Organizational Change." Organization Science, (18:5), pp.832-848.

[46] Wang, X., Dugan, R., & Sojka, J. (2013). CRM Systems with Social Networking Capabilities: The Value of Incorporating a CRM 2.0 System in Sales/Marketing Education. *Marketing Education Review*, *23*(3), 241–250. https://doi.org/10.2753/mer1052-8008230303

[47] Wilson, H., Daniel, E., & McDonald, M. (2002). Factors for Success in Customer Relationship Management (CRM) Systems. *Journal of Marketing Management*, *18*(1–2), 193–219. https://doi.org/10.1362/0267257022775918

[48] Wu, S.-I., & Hung, J.-M. (2008). A performance evaluation model of CRM on non-profit organisations. *Total Quality Management & Business Excellence*, *19*(4), 321–342. https://doi.org/10.1080/14783360701591978

[49] Zaby, C., & Wilde, K. D. (2017). Intelligent Business Processes in CRM. *Business & Information Systems Engineering*, *60*(4), 289–304. https://doi.org/10.1007/s12599-017-0480-6

Appendix

A. Interview questions semi-structured interview

The interview questions are presented below:

- How would you describe your current function?
- What is your experience in the field of data analysis?
- Which techniques/programmes do you use for data analysis and why?
- Which techniques/programs would you recommend for data analysis using internal and external data?
- Do you have examples of projects where you worked with an integration between internal and external data if yes, please elaborate?
- What are the pitfalls when performing an analysis between internal and external data?
- How could these be overcome in your opinion?
- Do you think the use of external data could improve processes within an organization if yes, how?

A1. Results interview 1

How would you describe your current function?

I am a consultant for incident management, where I give advice on software management and I also program myself. My specialty is in the field of incident management, which is about the accidents that happen on the road. I help with projects from Rijkswaterstaat and regional traffic management together with a university. Here I help with the rules that should be followed if a car has broken down on the road and help needs to come. I also used to work at the University of Valencia where my specialization was geoinformatics.

What is your experience in the field of data analysis?

All my projects can be seen as data-driven projects. At Rijkswaterstaat but also the other projects, data analysis was important. Here the quality of the data was of utmost importance for a good analysis.

How can good data quality be assured?

Data cleaning is very important. Most of the data is inserted by humans, this means that it must be made easy for them to insert the right data. By making use of standardization, the level of errors in the data will decrease. Teaching your employees the right way of working is also important if the insertion fields can not be standardized. A little difference in how people insert data in the system can have a huge influence on the outcome of the data analysis. Do not give the employee too many options, that way the level of errors will reduce. Also inform the employees why they have to insert the data that they insert if they are aware of the function of the data, they will act more carefully with it. Most employees do not like it when people keep looking at what they are doing, but they have to know how important the correct insertion of data is to the company.

Do you have a preference when it comes to data cleaning between manual and automatic systems?

People should dare to use automatic systems combined with some manual tweaking afterward. Manual transformation of the data is important, when this is done it should be recorded who and when transformed the data, to ensure that the data is not incorrectly transformed.

Which techniques/programs do you use for data analysis and why?

I use mostly special databases, with time/space analysis and a rule engine that on the fly the inserted data checks. This was chosen because I was already familiar with it and I did not want a vendor lock, with which I would be locked to one vendor, other options were still to new and had flaws so I chose this one. There are more fancy ways of doing it which are way more expensive, but the current way is cheaper and works. When searching for the right system I always say: If you do not know what you are looking for, you are not going to find it. That is a problem at a lot of companies nowadays, they buy something without knowing what they actually need. If you know what you are searching for in your data, you know what data needs to be saved and based on that buy the right systems for it. The technique that you want to use should easily fit what you need. The technique that I use is quite old, but still works for what I need. We also make use of a lot of open-source systems, these are also good and cheap solutions. Another useful technique is linked data, but currently, this is not yet very popular are companies.

Which techniques/programs would you recommend for data analysis using internal and external data?

The previous answered technique of mostly special databases, with time/space analysis and a rule engine that on the fly the inserted data checks works for this kind of data analysis. We do make use in our projects of a combination of internal and external data and this technique does work.

Do you have examples of projects where you worked with an integration between internal and external data if yes, please elaborate?

At a project at Rijkswaterstaat, the goal was to find out that an incident had happened on the highway sooner. It happened too often that another institution found out about it earlier. This is why we performed an analysis of our own data combined with social media data and Flitsmeister data as an external source. This combination made it possible to be aware of the incident sooner.

What are the pitfalls when performing an analysis between internal and external data?

The first one is that the quality of the data should be good. There should be a healthy distrust towards data that is inserted manually. A lot of data that comes in is bad in my experience and has to be adjusted or thrown away, roughly about 50 percent. Some data might not be complete, make sure that all the data that is required for the analysis is saved.

Saving data just because it is possible to save the data should not be done. Research has to be done before the analysis part to know what kind of data is needed, otherwise, the database will be flooded with too much unnecessary data, which makes the system slower and less clear. There should be a logical reason behind all the data that is gathered. It also reduces the workload on employees while inserting data.

The frequency of data saving can also be too much. Do not insert data more times than needed or the database will be unusable.

How could these be overcome in your opinion?

Assess the data on whether it is a good addition to the database or not and only use what is needed and always check data that is inserted manually, it is okay to be a little distrustful towards manual inserted data in order to increase the quality of the database.

Do you think the use of external data could improve processes within an organization if yes, how?

Yes, I think so, as long as there is commercial interest for the company to act upon the advice of the data analysis. If the internal/external analysis shows that something can make the company more profitable or improve it in another way, then the analysis is highly usable.

To improve processes using internal/external data, it is important to perform research before the actual analysis to find out where the focus of the research should be.

A2. Results interview 2

How would you describe your current function?

I am a business consultant. My primary role is to answer questions of clients about their data and systems provided by us. The focus is mostly on the processes within companies. I help them by providing them with the right software package or solutions to questions regarding the packages. I started three years ago at Lumen BS.

What is your experience in the field of data analysis?

I have a lot of experience from my study, data analysis was one of my courses. Currently at Lumen I also work with it a lot, the difference between data analysis between my study and at Lumen is that at Lumen we use premade tools instead of building everything from scratch and optimizing the loading times. At my current employee we try to make the data as clear as possible for our clients and let the data work for us, instead of only being used for reports afterward. For example, we try to find out how processes can be improved and the earnings of a company can be increased. I also choose to work more actively on using CRM systems for data analysis instead of using it as a data warehouse. When a client comes with certain assumptions we can test to see if this is indeed the case, with the help of visualizations in a dashboard.

Do you also help the client with finding out what they want to know for the analysis?

In most cases the client comes to us with a problem that they cannot figure out with their current tools. The client asks a specific question, but we do notice that the questions of clients are fairly similar, so we are mostly prepared for their questions. Once we have provided them with a system and have trained them they can tweak the system themselves to answer their questions.

Which techniques/programs do you use for data analysis and why?

We use Power BI for the analysis. There are also possibilities using Azure, but we have not had the time to get to know this program better. We chose to use Power BI because we work together with Microsoft and they advise this product. For clients, the price is important and Power BI is quite cheap to use for CRM and ERP analysis. In addition it is not hard to learn. It shows clearly the visualizations, the tables with data and the relationships between the tables.

How do you determine the relationships between the data sets?

Before Corona, we went to the client to find out where the connections between the multiple table were. Now we do this in online meetings, here we try to find out how the data should be interpreted and where the connections should be.

What are important properties for data to be usable for data analysis?

It is important to keep your data structured and clean. To do this the fields that the employees need to fill in should be standardized and the data should be normalized, this means that data should not be unnecessary in multiple tables. Ideally, data only occurs in one table. If there is a field with not a lot of possible answers we like to use a dropdown, this reduces the chance of a wrong insertion of data. Another important aspect of the data is, that all the included parties should interpret the data the same way. Amsterdam could be in the Netherlands but also in America, so for this example longitude and latitude should be used. This can also be the case for other types of data so keep that in mind.

Which techniques can be used to create prediction models based on the available data sets?

Trend analysis can be used in Power BI to generate a prediction model. Here the programs try to predict the future based on previous data, but not in the smartest way. To get a better prediction I would advise using an AI connection from Azure that produces more advanced prediction models. There are multiple options with Azure, that can be used and give a good prediction.

Which techniques/programs would you recommend for data analysis using internal and external data?

Power Bi can be useful for this as long as the database is available and does not need to request the external data for every insertion in the data set. Requesting live data from external data sources is difficult because Power BI is mostly used for data that is already in the system. But if these pitfalls are overcome I would use Power BI or perform a query analysis myself. Power Bi is However easier to learn than making a query analysis so that is why Power BI would be my recommendation.

Do you have examples of projects where you worked with an integration between internal and external data if yes, please elaborate?

We have not done it before at Lumen, but I did do it in my private time for a project. For this analysis, I did use Power BI.

What are the pitfalls when performing an analysis between internal and external data?

What I find most important is to sit together with the client at the end of the project to see what they think of the findings. In most cases the client did not expect the outcome that I show them. This means that he has learned something new about his company, but in most cases there is still a mistake in the dashboard, due to filters missing or data that should be interpreted differently. That is why it is important to sit together with the client and improve the analysis.

How could these be overcome in your opinion?

Stay in contact with your client, during the analysis but especially after the analysis. Always verify with the client that the found conclusions are true. Show not only the end results, but also show how you did get at these results.

Do you think the use of external data could improve processes within an organization if yes, how?

The usage of resources is most often something that can be improved. Where each resource must be can be analyzed with the help of the internal/external analysis, so it could improve processes of organizations. If you could make a pro-active planning based on the internal/external analysis then this could benefit the organization a lot.

B. Questionnaire results

Results of Questionnaire

Performance expectancy					
I Would find the approach useful in my job				X	
Using the approach enables me to accomplish taks more quickly			X		
Using the approach increases my productivity				X	
Effort Expectancy					
My interaction with the approach would be clear and understandable				X	
It would be easy for me to become skillful at using the approach				X	
I would find the approach easy to use				X	
Attitude Towards Using Technology					
Using the approach is a good idea				X	
The approach makes work more interesting					X
Working with the approach is fun			X		
I would like to work with the approach				X	
Facilitating Conditions					
I have the resources necessary to use the approach			X		
I have the knowledge necessary to use the approach			X		
The approach is not compatible with other approaches I use		X			
	Strongly disagree	**Disagree**	**Neutral**	**Agree**	**Strongly agree**
Self-Efficacy					
I would be capable of using the approach if....					
There was no one around to tell me what to do as I go		X			
I could call someone for help if I got stuck			X		
I had a lot of time to complete the job for which the apprach is designed				X	
Anxiety					
I feel apprehensive about using the approach		X			
I hestitate to use the approach for fear of making mistakes I cannot correct		X			
The apprach is somewhat intimidating to me		X			

Figure 42 Results Questionnaire 1

Performance expectancy					
I Would find the approach useful in my job				X	
Using the approach enables me to accomplish taks more quickly				X	
Using the approach increases my productivity				X	
Effort Expectancy					
My interaction with the approach would be clear and understandable			X		
It would be easy for me to become skillful at using the approach				X	
I would find the approach easy to use				X	
Attitude Towards Using Technology					
Using the approach is a good idea				X	
The approach makes work more interesting			X		
Working with the approach is fun		X			
I would like to work with the approach				X	
Facilitating Conditions					
I have the resources necessary to use the approach				X	
I have the knowledge necessary to use the approach			X		
The approach is not compatible with other approaches I use		X			
	Strongly disagree	**Disagree**	**Neutral**	**Agree**	**Strongly agree**
Self-Efficacy					
I would be capable of using the approach if....	I have sufficient knowledge of the customer case and available sources				
There was no one around to tell me what to do as I go			X		
I could call someone for help if I got stuck				X	
I had a lot of time to complete the job for which the apprach is designed		X			
Anxiety					
I feel apprehensive about using the approach		X			
I hestitate to use the approach for fear of making mistakes I cannot corre	X				
The apprach is somewhat intimidating to me		X			

Figure 43 Results Questionnaire 2

Performance expectancy					
I Would find the approach useful in my job			X		
Using the approach enables me to accomplish taks more quickly				X	
Using the approach increases my productivity				X	
Effort Expectancy					
My interaction with the approach would be clear and understandable			X		
It would be easy for me to become skillful at using the approach			X		
I would find the approach easy to use			X		
Attitude Towards Using Technology					
Using the approach is a good idea				X	
The approach makes work more interesting			X		
Working with the approach is fun			X		
I would like to work with the approach				X	
Facilitating Conditions					
I have the resources necessary to use the approach			X		
I have the knowledge necessary to use the approach			X		
The approach is not compatible with other approaches I use			X		
	Strongly disagree	**Disagree**	**Neutral**	**Agree**	**Strongly agree**
Self-Efficacy					
I would be capable of using the approach if....					
There was no one around to tell me what to do as I go			X		
I could call someone for help if I got stuck				X	
I had a lot of time to complete the job for which the apprach is designed			X		
Anxiety					
I feel apprehensive about using the approach			X		
I hestitate to use the approach for fear of making mistakes I cannot correct			X		
The apprach is somewhat intimidating to me			X		

Figure 44 Results Questionnaire 3

Performance expectancy					
I Would find the approach useful in my job					X
Using the approach enables me to accomplish taks more quickly				X	
Using the approach increases my productivity				X	
Effort Expectancy					
My interaction with the approach would be clear and understandable					X
It would be easy for me to become skillful at using the approach				X	
I would find the approach easy to use			X		
Attitude Towards Using Technology					
Using the approach is a good idea					X
The approach makes work more interesting				X	
Working with the approach is fun				X	
I would like to work with the approach				X	
Facilitating Conditions					
I have the resources necessary to use the approach			X		
I have the knowledge necessary to use the approach			X		
The approach is not compatible with other approaches I use		X			
	Strongly disagree	**Disagree**	**Neutral**	**Agree**	**Strongly agree**
Self-Efficacy					
I would be capable of using the approach if....					
There was no one around to tell me what to do as I go			X		
I could call someone for help if I got stuck		X			
I had a lot of time to complete the job for which the apprach is designed				X	
Anxiety					
I feel apprehensive about using the approach		X			
I hestitate to use the approach for fear of making mistakes I cannot correct		X			
The apprach is somewhat intimidating to me		X			

Figure 45 Results Questionnaire 4

Performance expectancy					
I Would find the approach useful in my job				X	
Using the approach enables me to accomplish taks more quickly		X			
Using the approach increases my productivity			X		
Effort Expectancy					
My interaction with the approach would be clear and understandable			X		
It would be easy for me to become skillful at using the approach				X	
I would find the approach easy to use				X	
Attitude Towards Using Technology					
Using the approach is a good idea			X	X	
The approach makes work more interesting		X			
Working with the approach is fun			X		
I would like to work with the approach				X	
Facilitating Conditions					
I have the resources necessary to use the approach				X	
I have the knowledge necessary to use the approach				X	
The approach is not compatible with other approaches I use	X				
	Strongly disagree	**Disagree**	**Neutral**	**Agree**	**Strongly agree**
Self-Efficacy					
I would be capable of using the approach if....					
There was no one around to tell me what to do as I go				X	
I could call someone for help if I got stuck					X
I had a lot of time to complete the job for which the apprach is designed		X			
Anxiety					
I feel apprehensive about using the approach	X				
I hestitate to use the approach for fear of making mistakes I cannot corre	X				
The apprach is somewhat intimidating to me			X		

Figure 46 Results Questionnaire 5