INTERNSHIP REPORT

Data Vigilance: Advancing Integrity Assurance and Automated Excellence

Date completed: 18-06-2024.

Author: Ramy Alashabi

Student number: 3765529

Version: 0.17

Status: Final

18 JUNI 2024

TWENTYNEXT

Eindhoven

GRADUATION-INTERNSHIP Thesis BACHELOR-ICT

FONTYS UNIVERSITY OF APPLIED SCIENCES

Student:	
Family name , initials:	**Alashabi, R**
Student number:	**3765529**
project period: (from – till)	**19-02-2024/ 12-07-2024**
Company:	
Name company/institution:	**Twentynext**
Department:	**Data**
Address:	**Kennedyplein 246, 5611 ZT Eindhoven**
Company mentor:	
Family name, initials:	**Maasakkers, M**
Position:	**Founder**
University teacher:	
Family name , initials:	**Lycklama, R**
Final portfolio:	
Title:	Data Vigilance: Advancing Integrity Assurance and Automated Excellence
Date:	**18-06-2024**

Approved and signed by the company mentor: Maikel Maasakkers

Date: 18-06-2024

Signature:

Inhoudsopgave

Table of figures 4
Preface 5
Document history 6
List of terms & abbreviations 7
Management summary 8
1. Introduction 9
1.1. Reader's guide 9
2. Context and background 10
2.1. About the client 10
2.2. Current situation 10
2.3. Problem analysis 12
3. Project statement 13
3.1. Project goal 13
3.2. Requirements 13
3.3. Research questions 13
3.4. Research approach 14
3.5. Research methods used 14
4. Research 15
4.1. Sub question 1 15
4.1.1. Results 16
4.1.2. Conclusion 20
4.2. Sub question 2 21
4.2.1. Results 21
4.2.2. Conclusion 25
4.3. Sub question 3 26
4.3.1. Results 26
4.3.2. Conclusion 39
5. Conclusion and recommendations 40
5.1. Conclusion 40
5.2. Recommendations 41
Assignment evaluation 42

Personal reflection 45
Bibliografie 48
Appendices A: project plan 53
Appendices B: research document 79

Table of figures

Figure 1 Introduction on the Medallion architecture, layers and their functions. 11
Figure 2 Methods used for the first sub-question. 15
Figure 3 All aspects that data governance framework covers 17
Figure 4 Screenshot of the best practices made in the research document. 18
Figure 5 Screenshot of the testcases made in the research document. 20
Figure 6 Methods used for the 2nd sub-question. 21
Figure 7 Screenshot of the metrics made in the protocol document 22
Figure 8 Screenshot of excel document. 23
Figure 9 Screenshot of the robust Jupyter code. 23
Figure 10 Screenshot of error found due to codes made. 24
Figure 11 Screenshot of excel document documenting patterns and new rules/metrics. 24
Figure 12 Methods used for the 3rd sub-question. 26
Figure 13 Screenshot of research document showing User requirements. 27
Figure 14 Screenshot of research document showing Functional requirements. 27
Figure 15 Screenshot of research document showing Non-functional requirements. 27
Figure 16 Showing how to connect Databricks and Power BI. 28
Figure 17 Screenshot shows how python codes were translated into DAX codes. 28
Figure 18 Showing profiling function in Power BI. 28
Figure 19 Showing the semantic model and relations. 29
Figure 20 Shows the columns made from python. 29
Figure 21 Shows the home page and navigation of the dashboard 30
Figure 22 Shows the first visuals made for the BSN test. 30
Figure 23 Shows the wishes of the stakeholder. 32
Figure 24 Shows the new iteration of the wished visual. 33
Figure 25 First iteration of the drillthrough function based on the requirements 34
Figure 26 Showing the number of errors in separate visuals. 34
Figure 27 Showing the Filter button. 34
Figure 28 Showing the power of bookmarks to save space on the dashboard. 35
Figure 29 Showing new ways to display filtered values. 35
Figure 30 Showing the struggles during the pivoting. 36
Figure 31 Showing the wanted results without the need of pivoting 36
Figure 32 Showing that empty rows had to include a string "Empty" for the error to show. 37
Figure 33 Showing the main page of the final iteration of the dashboard 37
Figure 34 Showing the drill through table format. 38
Figure 35 Showing the final iteration of the drill through table. 38
Figure 36 Showing the merge of the tables. 38
Figure 37 Showing how to disable dataload. 39

Preface

This document is a report written for my final semester/graduation project at Fontys University of Applied Sciences. The project, titled "Data Vigilance: Advancing Integrity Assurance and Automated Excellence," was assigned to me by a company called Twentynext. I was introduced to the company in the final week of submitting my proposal through a recruiter who had contacted me on LinkedIn a year earlier. After meeting with Maikel Maasakkers and Martijn van Grieken, I secured an internship and began this project. From February 2024 to June 2024, I have dedicated my efforts to researching, reporting, and realizing the goals of this project.

During my previous semesters, I realized that my documentation skills were not as strong as they should be. Often working in groups, I would stay in my comfort zone by taking on the coding tasks while leaving the documentation to my teammates. This internship project provided an invaluable opportunity to work independently in a real-world environment, where theoretical knowledge meets practical application. Stepping out of my comfort zone allowed me to gain experience in collaborating within an organization, significantly improving my report-writing skills, and further enhancing my coding abilities. Additionally, this project gave me the chance to apply the knowledge and skills I had previously taught myself, integrating them effectively into my work. This experience has been instrumental in my professional development, equipping me with a well-rounded skill set essential for my future career. During this internship, I have greatly appreciated the support and guidance provided by my work mentor, Maikel Maasakkers, and my co-workers, Ni Kang and Linda Bujitu. Their patience and willingness to help were invaluable, especially when I had numerous questions and needed repeated explanations. Their dedication to assisting me in my learning journey was truly remarkable. I am also grateful to my school mentor, Rink Lycklama, who taught me to think visually and guided me through the writing of my project and the structuring of its scope. His insights were crucial to the successful completion of this project.

I would like to extend my heartfelt thanks to my mentors and everyone who provided feedback and assistance in structuring my work and documents throughout this project. Special thanks go to Martijn van Grieken and Maikel Maasakkers for giving me the opportunity to work on and realize this project at Twentynext within a short time frame.

Additionally, I would like to express my sincere gratitude to Ni Kang for offering her help and feedback, and to Elles Dillen and Jasper de Kort for their motivation and warm welcome. Their support made a significant difference in my experience and contributed greatly to my personal and professional growth during this internship.

May the force be with you all,

Ramy Alashabi
Eindhoven, 17th June 2024

Document history

Revision

Version	Status	Date	Changes
0.1	Draft	13-04-2024	Finished ch1-2
0.2	Draft	15-04-2024	Finished ch3-4.1
0.3	Draft	25-04-2024	Working the feedback from Rink
0.4	Draft	06-05-2024	Feedback from Maikel
0.5	Draft	13-05-2024	Feedback from NI
0.6	Draft	17-05-2024	Feedback from Ni on research
0.7	Draft	27-05-2024	Added a new method to the 3rd question.
0.8	Draft	30-05-2024	Finished 3rd iteration on 3rd question
0.9	Draft	31-05-2024	Finished conclusion of 3rd question
0.10	Draft	03-06-2024	Finished conclusion and recommendation
0.11	Draft	04-06-2024	Implementing Rinks Feedback
0.12	Draft	10-06-2024	Wrote evaluation
0.13	Draft	11-06-2024	Wrote the reflection
0.14	Draft	12-06-2024	fixed the table of figures and tables
0.15	Draft	13-06-2024	Fixed the references and appendices
0.16	Draft	17-06-2024	Final edits from work mentor, wrote the management summary and preface.
0.17	Final	18-06-2024	Working on last feedback from work mentor.

Distribution

This document is distributed to:

Version	Date	Name	Function
0.2	16-04-2024	Rink	School Mentor
0.3	26-04-2024	Maikel	Company mentor
0.4	06-05-2024	Rink and Ni	School mentor and co-worker
0.5	13-05-2024	Ni	Co-worker
0.6	21-05-2024	Ni and Rink	Co-worker and mentor at school
0.10	03-06-2024	Rink	School mentor
0.15	13-06-2024	Maikel	Company mentor

Approval

This document requires following approvals:

Version	Date approval	Name	Function	Signed

List of terms & abbreviations

Terms and abbreviations	Meaning
ETL	Extract, Transform, Load
EDA	Exploratory data analysis
EHR	Electronic Health Record
HIS	Health information system
DOD	Date of death
DOB	Date of birth
i.e.,	used to explain exactly what the previous thing that you have mentioned means (from Latin ‘idest’)
vividly	in a way that produces very clear pictures in your mind
albeit	although

Management summary

The name of this project is “Data Vigilance: Advancing Integrity Assurance and Automated Excellence”, it was commenced on the 19th of February 2024 till the 18th of June 2024, for Twentynext. Twentynext is a business intelligence company that is working for a nursing home called “AxionContinu.” The project aimed to address critical challenges faced by AxionContinu in maintaining data integrity and quality throughout the ETL process within nursing home data systems. Through a structured approach, encompassing research, protocol development, and dashboard creation, significant strides were made towards achieving this goal. The project started with a main question “How can a methodical instrument be designed that effectively measures data integrity, incorporating flexible parameters aligned with business requirements, while ensuring consistency and reliability in the assessment process?” Based on that 3 sub questions were evolved to help in answering the main questions, and they are as follow:

1. How can nursing home data systems ensures the integrity of their data (complete, accurate and consistent) through the development of a protocol?
2. What metrics or code of validation measures should be considered for ensuring data integrity, and can they be tailored for diverse solution requirements?
3. How can an integration into the IT landscape be incorporated, while measuring the dimension of data integrity and employing suitable visualization in a monitoring dashboard?

Through the first sub question, it was possible to identify and define data quality, focused on exploring existing frameworks, metrics, and protocols related to data integrity. Insights gleaned from literature studies, document analyses, best practices identification, and interviews with industry experts informed the development of a robust protocol aimed at ensuring data accuracy, completeness, consistency, timeliness, validity, and uniqueness. The second sub question was focused on developing a protocol with metrics that validates the current and future data. Using industry standards such as ISO 8000 and insights from interviews with data scientists, data engineers, and IT professionals, a comprehensive protocol was devised. This protocol outlined clear rules for maintaining high-quality data throughout its lifecycle, thereby facilitating better decision-making and operational efficiency. The third sub question was focused on the development of the dashboard and collection of the data in the IT- landscape using Power BI. The dashboard evolved to effectively visualize errors, data profiling, and provide drill-down functionalities as per stakeholder requirements. In conclusion, the project addressed the primary research question successfully by developing a methodical instrument for measuring data integrity. True that the project achieved a significant milestone. However, further development, testing and feedback collecting is recommended to ensure continuous improvement and alignment with the evolving business needs. This document shows every step taken from the beginning of the project till the end, it represents the approach, what is expected of the approach and the results of the approach. Based on this document, this project can be recreated and can be improved on in the future.

1. Introduction

This document serves to provide an overview of the "Data Health in Healthcare" project conducted by the business intelligence firm "Twentynext" on behalf of the nursing home "AxionContinu." AxionContinu sought Twentynext's assistance in consolidating multiple data sources into a unified data and reporting platform, with Twentynext prioritizing the maintenance of data integrity throughout the consolidation process. Therefore, the primary objective of this document is to outline the identified problem, detail the research conducted to address it, and present the prototype developed as a solution.

1.1. Reader's guide

This document contains 7 chapters as can be seen below:

1. The first chapter is an introduction about this document and the objective of the document.
2. The second chapter serves the purpose of introducing the current problem and the company.
3. The third chapter is the project statement chapter, where the goals, research questions and methods used.
4. The fourth chapter is the research chapter. Where the research, preliminary research and the implementation will be reported.
5. The fifth chapter will elaborate the recommendation and conclusion of the project.
6. The sixth chapter serves the purpose of an evaluation of the project.
7. The seventh chapter serves the purpose of a personal reflection.

The rest are the sources and the appendices.

***NOTE*: Appendices B Research document contains more appendices that are not included in the thesis (including the interviews, extra screenshots etc.), as that would increase the word count of the document, due to repetition.**

2. Context and background

2.1. About the client

This chapter provides an overview of Twentynext, a company specializing in data science and artificial intelligence and consists of 19 employees. Twentynext helps organizations make better decisions by extracting value from data (Twentynext, 2024), such as retailers, health organization and supermarkets. They assist clients in setting up and managing complex data projects, offering expertise and support. One of Twentynext's clients is AxionContinu, a nursing home organization based in Utrecht, founded in 2013 (AxionContinu, 2023). AxionContinu sought Twentynext's help with an IT and data governance issue, which will be discussed further in the following chapter.

2.2. Current situation

The purpose of this chapter is to describe the current situation of the Twentynext's project regarding AxionContinu and how it is linked to this internship project. AxionContinu, a client of Twentynext, seeks assistance to address an issue in their current process. Currently, data and dashboards are sourced from four different data warehouse, leading to confusion and inconsistencies. These 4 warehouses are:

1. Infent
2. Cognos
3. OGD
4. mijnCaress

Infent typically operates within the "InTouch" dashboard, showcasing financial data, sick leaves, and is communicating with mijnCaress. While mijnCaress contains operational information about the employee, patients, and health records, Infent warehouse contains the financial information of those employees and patients.

Cognos, OGD, and mijnCaress collaborate on mijnCaress data including patient information, schedules, and more. This dependent approach by AxionContinu led to false data, especially when the patient is registered differently in a different warehouse, hence not trusted by the employees/nurses. This hinders AxionContinu's access to comprehensive information essential for their tasks and operations, such as not being able to bill a patient due to name inconsistencies within the warehouses. Alongside high maintenance costs of the warehouses for management department, which, according to the stakeholder, could cost around 500,000 USD minimum per year (i.e., IT support 200K, Suppliers 200K, Licenses and hardware 100K, etc.) (Whiting, 2024), and the inability to execute data-driven work effectively due to inconsistencies. Additionally, most organizations implement a failover system to mitigate risks in case the primary system fails (Rajeshsetlem., 2024). The key

difference between a failover and a backup is that a backup is a copy of data that can be restored when needed, while a failover automatically switches to another system when the primary system fails. Two critical metrics in these systems are the Recovery Point Objective (RPO) and the Recovery Time Objective (RTO) (Singh, 2021). RPO indicates how much data loss is acceptable, while RTO specifies how quickly the system should be restored after a failure. These metrics are essential for effective disaster recovery planning and minimizing business disruptions. Nevertheless, to reducing the maintenance cost, Twentynext proposed a solution that involves streamlining data sources from four to one. Progress has been made by collecting data from Software as a Service (SaaS) platforms and depositing it into a Datalake as an initial step towards consolidating data sources.

After the data collection phase, which was done in azure data factory (ADF), Twentynext utilizes Databricks for ETL operations, adhering to the Medallion Lakehouse architecture (Databricks, 2024) shown in Figure 1.

This architecture, represented by bronze, silver, and gold layers, serves to systematically enhance data quality as it progresses through each layer. Each layer within this architecture fulfils a unique function, as illustrated below.

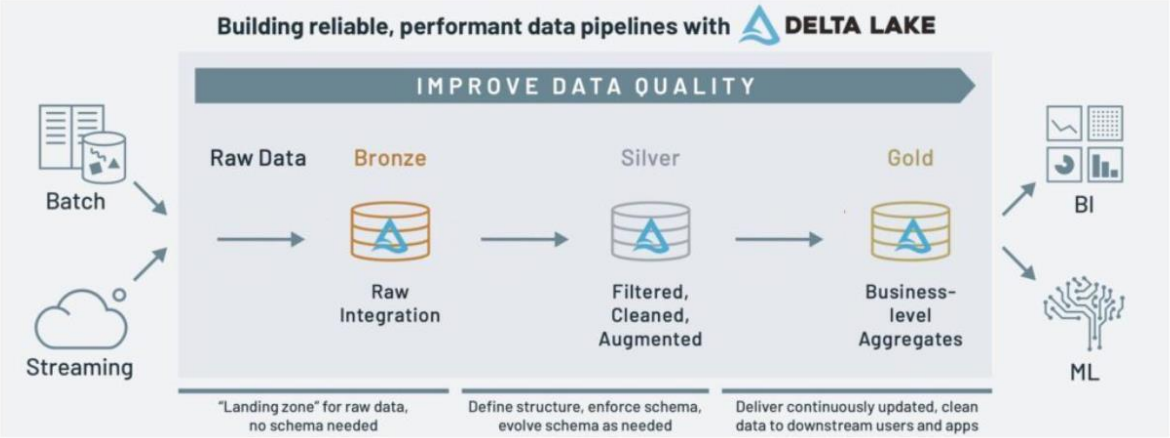


Figure 1 Introduction on the Medallion architecture, layers and their functions.

Within the medallion architecture, bronze involves raw data collection, while silver focuses on structuring and cleansing the data. Gold entails constructing data models with factual and dimensional tables, representing business-level aggregates. Currently, the project is transitioning between the bronze and silver layers, and no actions have been taken regarding data quality.

According to a colleague, a proper transition is crucial to ensure the data integrity of the next steps within the silver and gold layers (Versantvoort, 2024).

2.3. Problem analysis

During a discussion with one of the stakeholders, it was revealed that AxionContinu utilizes a ECD (Electronic client dossiers) (Van Der Burg, 2023) system to manage patient records within mijnCaress. This system is divided into patients' data, personnel data, and financial data, with manual input being utilized. For instance, in cases of unknown gender or name, manual input such as a question mark or space is entered instead of "NAN" (Not A Number). However, this manual input is not monitored by AxionContinu, leading to errors such as registering the Date of Birth (DOB) after the Date of Death (DOD). Therefore, as part of the consolidation project, Twentynext decided to investigate the data integrity, identifying patterns, errors, and rules that could be discussed with the business side of AxionContinu. This would serve as advice on how to handle inconsistent data, such as dropping entries with a BSN exceeding 9 characters from the main database and validating them manually.

Main problems can be categorized as:

1- Introduces inconsistencies due to human input errors (i.e., DoD bigger than DOB) caused by manual data ingestion.
2- Lack/ absence of mechanisms of Data Monitoring and Validation undermines data integrity (i.e., No method/tool to visualize the number of errors per record).
3- Absence of clear rules for data integrity (i.e., BSN 11-proef) during ETL processes worsens data inconsistencies and inaccuracies, potentially contaminating the dataset (Segment, 2023) with unflagged rows, such as those with missing BSNs, delaying guidance delivery and decision-making processes. The lack of guidelines for ensuring data integrity during ETL processes, the use of test data in production systems, and missing data due to system conversions are the causes.

3. Project statement

3.1. Project goal

The goal of this project is to implement validation measures on existing data, establish accepted data integrity standards, and develop a dashboard for visualizing errors to enable and increase the efficiency of data engineers in the ETL process and data monitoring process at Twentynext for AxionContinu.

3.2. Requirements

Twentynext has stated 2 business requirements, and they are summarized as follows:

1. BR01: AxionContinu aims to fortify its data governance framework by implementing procedures to maintain data integrity and quality throughout the ETL process, ensuring that at least 1 key data quality metric (consistency) is tracked.
2. BR02: AxionContinu seeks to enhance its data integrity assurance practices by introducing proactive measures for integrity validation and continuous monitoring of the current data repository, ensuring that 95% of the wanted data columns undergo automated integrity validation and daily monitoring is conducted for critical tables.

In this project, the requirements are collected from all conducted interviews and research done, as in there was no specific interview that was made only for requirements collections, but mostly translated from the weekly 1-on 1 meetings with the mentor into system requirements and user requirements. If you want to know more about the system and user requirements, please check appendices A the project plan's first appendices and Chapter 4 and the research document Chapter 2 for all the requirements. However, more about the requirements will be elaborated below in chapter 4.3.

3.3. Research questions

These are the research questions that this project is addressing:

Main question

How can a methodical instrument be designed that effectively measures data integrity, incorporating flexible parameters aligned with business rules, while ensuring consistency and reliability in the assessment process?

Sub questions

a. How can nursing home data systems ensures the integrity of their data (complete, accurate and consistent) through the development of a protocol?
b. What metrics or code of validation measures should be taken into account for ensuring data integrity, and can they be tailored for diverse solution requirements?
c. How can an integration into the IT landscape be incorporated, while measuring the dimension of data integrity and employing suitable visualization in a monitoring dashboard?

3.4. Research approach

In this section, the problems will be approached using various methods, such as:

1. Identifying challenges in data reliability and integrity at Axion Continu.
2. Investigating patterns in data like manual data entries, to identify underlying trends and issues.
3. Reviewing existing literature and consulting experts.
4. Engaging stakeholders for insights and requirements.
5. Testing potential solutions through prototypes.
6. Documenting findings and recommendations for implementation.

These methods will help in understanding the issues and identifying possible solution approaches.

3.5. Research methods used

Research methods are activities, tactics and procedures used in the collection of data or proof for analysis with the purpose of revealing additional information or establish a vivid understanding of the topic. There are different types of research methods and for this project we are using the methods provided by Fontys (Bonestroo, 2018) and they can be found in the Appendices A. Nonetheless, next chapter will elaborate every method used per question.

4. Research

This chapter will outline the research sub-questions and provide a detailed explanation of the research methods employed, along with elaboration on the results obtained. For Further technical details about the research check the documentation of the research in Appendices B

4.1. Sub question 1

How can nursing home data systems ensures the integrity of their data (complete, accurate and consistent) through the development of a protocol?
The research question aims to explore the concept of data integrity and investigate existing frameworks that could inform the development of a protocol ensuring data integrity. To answer this question, the following was used.

Figure 2 Methods used for the first sub-question.

- Literature study: used to research previous work done on data integrity, including frameworks such as (BaSIL and Cobit).
- Document analysis: used to analyze available document for information that were provided by Axion to Twentynext.
- Best good and bad practices: Identifying effective and ineffective practices to inform the development of the protocol while keeping the literature study in mind.
- Interview: conducting three separate interviews to validate the effectiveness of these best practices in real-world applications and correct them based on new insights.
- Guideline: using ISO 8000 as a standard constraint for the protocol.

Note: The protocol and results of both methods (Guidelines and best, good, and bad practices) must be documented separately, as they exceed 1200 words each and can be found in the appendices of the research document.

Note: The literature study on the governance frameworks is based on the free (UNPAID) version that are available/scattered around/on the internet.

4.1.1. Results

Literature study

The literature study provided insights into data quality, data governance frameworks, the Data Management Body of Knowledge (DMBoK), and ISO 8000 methodologies. Here are the key findings that are useful and applicable for the project:

- Data Quality Overview:
  - Helped in understanding the importance of data accuracy, completeness, consistency, validity, timeliness, and uniqueness for effective decision-making in businesses (AI, 2022).
  - Useful for forming benchmarks and criteria for assessing and improving data quality within the project (Sheldon, 2024).
- Data Governance Frameworks:
  - Recognition of various data governance frameworks, each with its pros and cons (Incept Data Solutions, Inc. , 2023).
  - Helpful for evaluating and selecting a framework aligned with the project's specific needs and priorities (ICTinformatiecentrum, 2023).
- DMBoK Framework:
  - Recognition of the DMBoK framework as a holistic guide for managing data effectively within organizations, including Rabo bank (Rabobank, n.d.).
  - Provides comprehensive guidance and best practices for developing data management strategies (Incept Data Solutions, Inc. , 2023).
- ISO 8000 Methodologies:
  - Introduction to ISO 8000 as a set of rules aimed at improving data quality, standardization, efficiency, and risk mitigation across industries (ISO 8000, 2022).
  - Offers guidelines for ensuring data accuracy, consistency, and reliability, which can inform the project's data management practices (Benson, ISO 8000: A new International Standard for Data quality, by Peter Benson , 2020).

These findings are useful for the project as they provide insights and guidelines on establishing data management practices within the ten core knowledge areas that can be seen below and within this project, we are strictly looking at the data quality part.

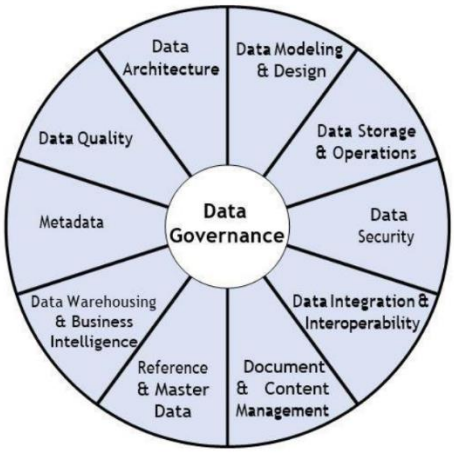


Figure 3 All aspects that data governance framework covers.

Some of these guidelines will be mentioned below in the Guidelines conformity method. However, for further details about the frameworks, please check the research document Appendices B.

Document analysis

Three documents were given by AxionContinu as documentations about their system, which were confusing and unstructured. All what is needed to be understood by the reader for now is that these documents were the only documentations available and were analyzed by the intern and they are called as follow:

- Align dumpstore documentatie
- Align infostore documentatie
- Align Kubus Documentatie infent infostore

Despite the challenges, the analysis of these documents provided some insights, albeit limited, into the system. The infostore document offered details on the length of each column in the tables, which was used to measure consistency. Similarly, the Kubus document indicated the types of data found within mijnCaress, such as financial or patient data. However, these documents primarily focused on the broader system rather than the specific sources of errors, contributing only marginally to understanding the data's characteristics. Even though the documentation added minimal value to the research, it was essential to document this process for future reference and learning. For further details about the analysis, please check the research document Appendices B.

Best, good and bad practices:

The protocol sets out clear rules to ensure data quality across different aspects like accuracy, completeness, consistency, timeliness, validity, and uniqueness. It provides guidelines for best, good, and bad practices in each area (derived from various sources such as ISO website and research articles per dimension), along with actions and examples of metrics for implementation that are derived from an interview which can be found in 8.5. Appendices E: Interview Linda

1. Accuracy: Making sure data is correct and comes from reliable sources.

2. Completeness: Ensuring all necessary information is present.

3. Consistency: Keeping data consistent across different sources.

4. Timeliness: Data constantly available and up to date when required.

5. Validity: Making sure data follows acceptable formats and rules.

6. Uniqueness: Ensuring each piece of data is unique.

By following these rules, data engineers can maintain high-quality data throughout its lifecycle, leading to better decision-making and operational efficiency. Below is an example of the practices.

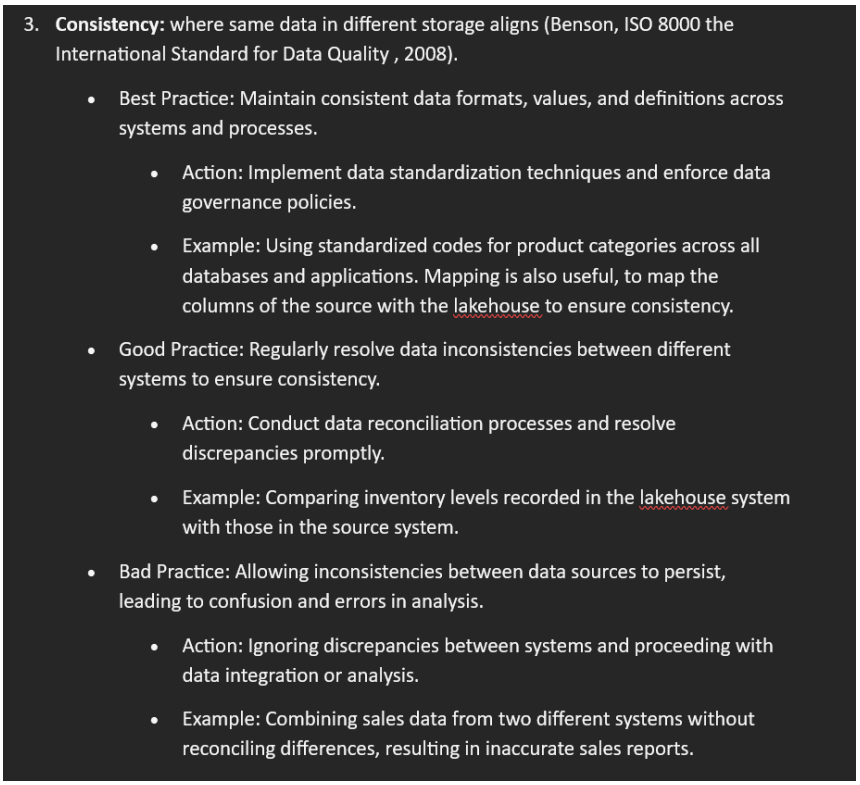

3. **Consistency:** where same data in different storage aligns (Benson, ISO 8000 the International Standard for Data Quality , 2008).
    - Best Practice: Maintain consistent data formats, values, and definitions across systems and processes.
        - Action: Implement data standardization techniques and enforce data governance policies.
        - Example: Using standardized codes for product categories across all databases and applications. Mapping is also useful, to map the columns of the source with the lakehouse to ensure consistency.
    - Good Practice: Regularly resolve data inconsistencies between different systems to ensure consistency.
        - Action: Conduct data reconciliation processes and resolve discrepancies promptly.
        - Example: Comparing inventory levels recorded in the lakehouse system with those in the source system.
    - Bad Practice: Allowing inconsistencies between data sources to persist, leading to confusion and errors in analysis.
        - Action: Ignoring discrepancies between systems and proceeding with data integration or analysis.
        - Example: Combining sales data from two different systems without reconciling differences, resulting in inaccurate sales reports.

Figure 4 Screenshot of the best practices made in the research document.

For further details about the practices, please check the research document Appendices B.

Interview

1. The first interview was with a data scientist. The purpose of the interview is to understand more about data quality and to get a perspective of a data scientist into the matter. During the interview we talked about how some machine learning techniques could be used to analysis the data quality and she elaborated more on how important it is to understand the data before doing anything. Interview can be found in appendices D.
2. The second interview was with a data engineer who was working by Rabobank. The purpose of the interview was to get information from a fellow data engineer on how to approach this problem and to get to know how it was done in Rabobank and to validate the best practices. This interview had an impact on the best practices and the protocol made prior to this method, as it makes it based on experience. The interview can be found in appendices E.
3. The Third interview was with an IT teacher. The purpose of that interview was to collect information on how he approached a similar situation and to validate the literature study. Additionally, a widespread discussion on various aspects of data management, including integrity, uniqueness, timestamping, hashing, primary keys, and business definitions. Key insights were shared regarding the importance of understanding business requirements, ensuring data integrity and uniqueness, and the need for collaboration between IT and business stakeholders. The interview can be found in appendices F.

Additionally, based on the interviews the protocol and the test cases were made.

Guideline conformity:

To ensure the credibility of the quality of the data, some guidelines and best practices are selected from ISO 8000 standards and DMBOK for the data quality dimensions (Richman, 2023). Per dimension, a testcase has been made with a hypothesized working result, to verify the compliance of the guidelines to the protocol (Linkedin, 2024). This is done to demonstrate the potential consequences of quality lapses, enabling reliable and trustworthy data while preventing inaccuracies and errors. Below is an example of this with some use case.

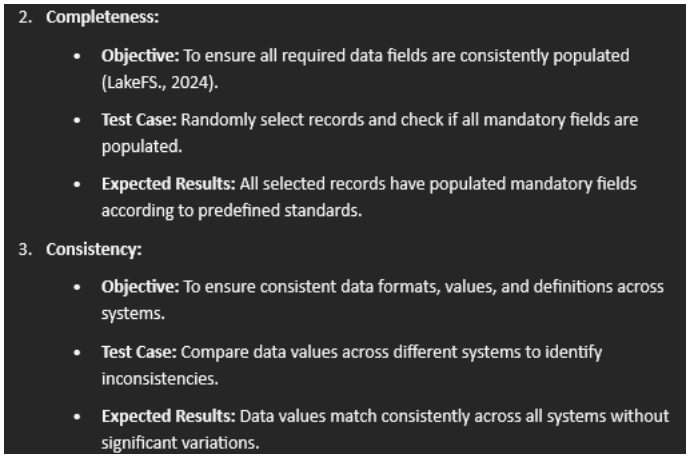

2. **Completeness:**
    - **Objective:** To ensure all required data fields are consistently populated (LakeFS., 2024).
    - **Test Case:** Randomly select records and check if all mandatory fields are populated.
    - **Expected Results:** All selected records have populated mandatory fields according to predefined standards.
3. **Consistency:**
    - **Objective:** To ensure consistent data formats, values, and definitions across systems.
    - **Test Case:** Compare data values across different systems to identify inconsistencies.
    - **Expected Results:** Data values match consistently across all systems without significant variations.

Figure 5 Screenshot of the testcases made in the research document.

The guidelines can be found in the research document Appendices B

4.1.2. Conclusion

This research tackled the creation of a protocol for better data integrity in nursing home systems, which contributed to the product. By researching, analyzing documents, gathering best practices, conducting interviews, and checking guidelines, a solid protocol was developed by the intern, that was shared to the engineer at Twentynext. This protocol, influenced by the findings of industry standards and expert advice, aims to improve data reliability and decision-making in nursing homes. As based on the interviews, test cases were developed and based on the frameworks, best practices were developed. Implementing it can lead to improved healthcare outcomes and efficient ETL processes for the engineer. Therefore, it can be concluded that the purpose of this sub question is achieved and can be proven with the document and results above.

4.2. Sub question 2

What metrics or code of validation measures should be taken into account for ensuring data integrity, and can they be tailored for business requirements?

The reason for this question is to research which measures can be used to validate and analyze the current data and to decide which dimension fits with what column. To answer this question, the following was used.

Figure 6 Methods used for the 2nd sub-question.

- Literature study: Used to research previous work done on data integrity and metrics that can be used to test the data (i.e. python codes/ rules).
- Ethical check: Used to determine which integrity dimensions apply to each table and column of the personal data stored in the company.
- Data quality check: used to analyze the current consistency and EDA of the current available data.
- Interview: used to confirm and review the metrics and protocol made if they have a potential impact.

4.2.1. Results

Literature study

Based on the research done on valuable metrics, keeping DMBOK and ISO 8000 in mind, an advised protocol document was made. The document includes the definitions of data integrity from the business perspective, Roles and Responsibilities, training and education and translation of the business requirements and metrics. The document can be found in Appendices C. Nonetheless, below is an example of the document.

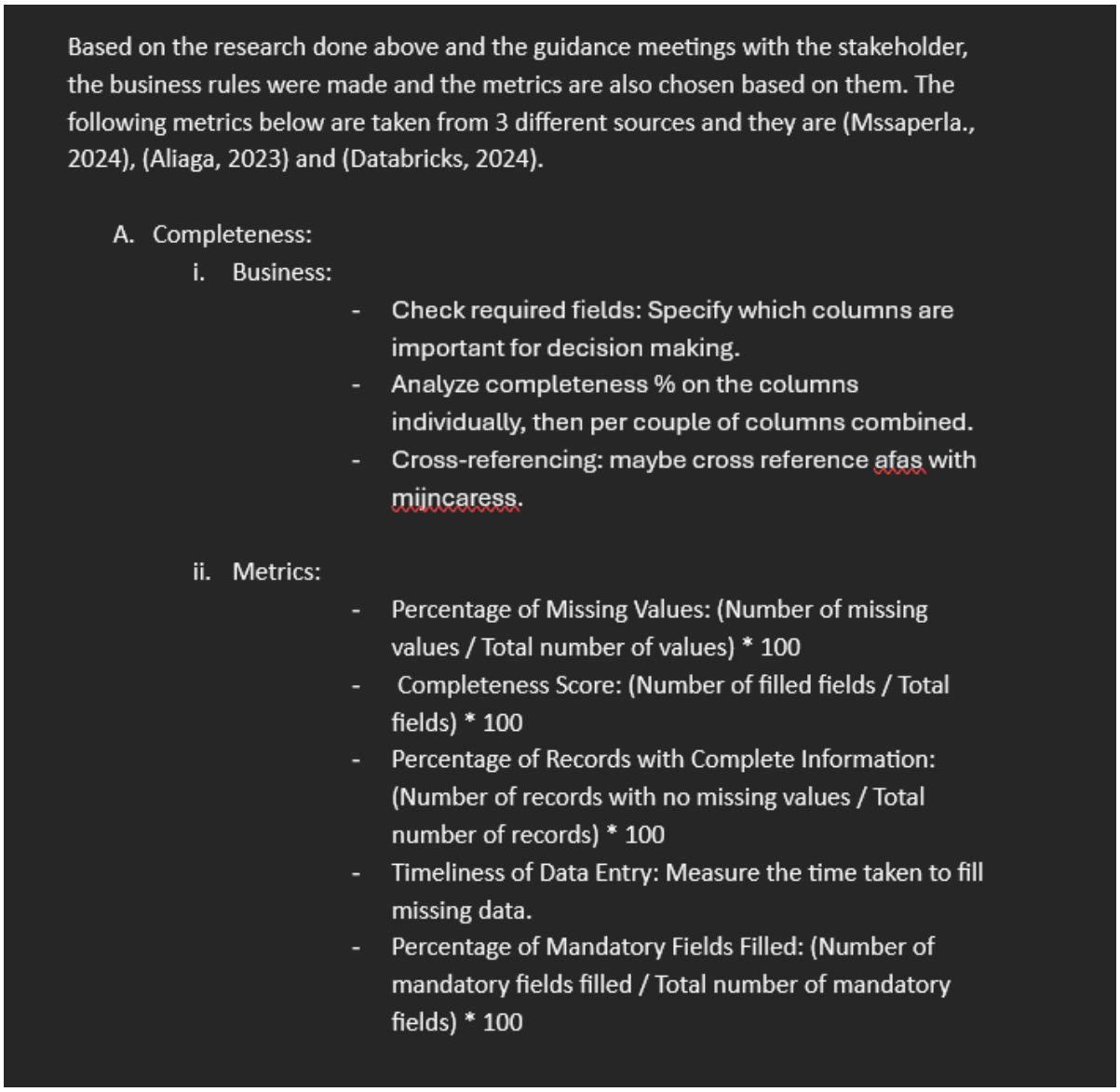


Figure 7 Screenshot of the metrics made in the protocol document.

Ethical check

The ethical check aimed to align the company's norms with the data being analyzed, which primarily consists of patient information such as date of birth, name, and addresses. The assessment of data integrity dimensions involved collaboration with a colleague to determine suitable fits for each column. During the check, it was realized that not all the dimensions of data integrity can be measured within these tables. As these are patients data and most of the errors are caused due to human input error (i.e., input of “?” or space instead of null or Nan). Therefore, an excel sheet was made to document the findings and what can be checked within consistency. A screenshot can be found below due to NDA the document cannot be uploaded to canvas.

Tabelnaam	Dimensions	KolomNaam	Voorbeeld Data	Data betekenis	Mogelijke Filters	Mogelijke kwaliteitscontroles
CoreTables						
Tbc	unqiue en consistant	IC		BK		Wisselende waarden
	complete en consistant	LongNm		Naam met voorletters	Naar UCASE of upper/down case	?
	uniqnue en consitant	Nr		patient number?		?
	constiancy, completeness	VrNm		Voornamen	Naar UCASE of upper/down case	Kan 1 of 2 of meer namen bevatten
	constiancy	CallNm		Roepnaam	Naar UCASE of upper/down case	ca
	consitancy	Vrv		Voor voegsel	Naar UCASE of upper/down case	c
	consistancy	AchtNm		Achter naam	Naar UCASE of upper/down case	?
	niks	Geslacht		Geslacht	Upper case	M of V --> in infostore Man/Vrouw
	accuracy, completeness	GebDat		Geboorte datum		Toekomst?
	accuracy	BirthPlace		Geboorte plaats	Naar UCASE of upper/down case	?
	accuracy	OvlDat		Overleidingsdatum		Na GebDat?, Toekomst?
	niks	Iusr		FK		Overal Null?
	check It out	Icountry		FK		?
	niks	Ilang		FK		Overal Null?
Tbc_Ad	unqinuess	ICAd		BK		?
	accuracy and complenteess	Straat		Straat naam	Naar UCASE of upper/down case	Bevat ".","?" en "1e/2e straatnamen" enz.
	accuracy	Hnr		Huisnummer		Huisnummers met 0 ervoor? Int van maken?
	accuray	HnrToe		Huisnummer toevoeging	Naar UCASE of upper/down case	?
	accuracy	Pc		Postcode	Upper case	Woon
	accuracy	Woonpl		Woonplaats	Naar UCASE of upper/down case	Bevat ".","?" en hoofdletters enz., Corrigieren op basis van postcode tabel? Omang met upercase/lowercase and speciale tekens
	check It out	ICountry		FK		Enige waarden die voorkomt of Null
	niks	IPcountry		FK		Enige waarden die voorkomt of Null
	unqinuess en constant	IC		FK naar Tbc		Wisselende waarden

Figure 8 Screenshot of excel document.

Data quality check:

This method is usually used for machine learning, however in this case it is a requirement from the business to check the quality of the current data. This process occurs in a Databricks notebook within the AxionContinu environment. In the notebook, the data is loaded and passed through a code snippet that verifies the consistency of all table columns. These code snippets are constructed to function seamlessly with either a new table or fresh data, ensuring adaptability across datasets, if the new data is renamed to “TBC”, as it can be seen below.

```
✓ 4/30/2024 (<1s)                                                    27

# consistancy range per column

def check_column_length(col_name):
    col_length= tbc[col_name].astype(str).apply(len)
    unique_lengths= col_length.unique()
    min_length=col_length.min()
    max_length= col_length.max()
    if len(unique_lengths)==1:
        print(f"ALL rows in column '{col_name}' have the same length: {unique_lengths[0]}")
    else:
        print(f"Not all rows in column  '{col_name}' have the same lengths.")
        print(f"lengths found: {unique_lengths}, miumum range:  {min_length}, max range: {max_length}")
        #diff_lengths= col_length[col_length.duplicated(keep=False)]
        #examples=tbc.loc[diff_lengths.index[:2], col_name]
        examples= tbc[col_name].iloc[[0,len(tbc)//2]].astype(str)
        highest_len=tbc[col_name].iloc[col_length.idxmax()]
        lowest_len=tbc[col_name].iloc[col_length.idxmin()]

        print(f"Value: {highest_len}, length: {len(str(highest_len))}")
        print(f"Value: {lowest_len}, length: {len(str(lowest_len))}")
        print(f"two examples where lengths differ:")
        for idx, value in examples.iteritems():
            print(f"Index: {idx}, Value: {value}, Count: {len(str(value))} ")


for column in tbc.columns:
    check_column_length(column)
```

Figure 9 Screenshot of the robust Jupyter code.

Moreover, discrepancies were identified and were documented in the excel sheet, such as instances where the date of birth is recorded after the date of death, as it can be seen below.

invalid_rows

	VR_NM	VRV	ACHT_NM	GESLACHT	GEB_DAT	BIRTH_PLACE	OVL_DAT	TEL1	BSN	SOFI_NR
21161		None		V	1946-05-07	Beni Badis	1921-09-27	(		

Figure 10 Screenshot of error found due to codes made.

Furthermore, a metric for the BSN was made to check the validity of the BSN (11-proef) and consistency. This metric is changed into a specification rule that Twentynext could use in advising their clients. Below is an example figure of the rules that were translated from the analysis.

Phone numbers	1. In nederland: 0NNNNNNNN (0+ 9 values) 2. Buitenland: 00NNNNNNNNNN (00+country code +9 values)
Bsn number	1. 9 numbers 2. Must pass 11 proef test.
Street name	Capital and lower cases, including punctuation marks (" ‘ ", " . ", "-").
House number	in the Netherlands this field should only contain integer numbers.

Figure 11 Screenshot of excel document documenting patterns and new rules/metrics.

For further elaboration please check the research/advise document Appendices B.

Interview

An interview was done with Linda who is a data engineer at twenty next and used to work with Rabo bank. The purpose of the interview was to validate the advice on the ETL process, and the metrics chosen in measuring the consistency of the data. The interview can be found in Appendices B within the appendices of the research/advice document. However, the interview can be summarized into the following:

- Good metrics and rules chosen for the ETL processes and asked for this to be shared with the others.
- In this project, consistency is the only valid dimension to look at, as to check completeness, accuracy, and uniqueness, we will either need the main data before bronze medallion or a different data set for cross reference.

4.2.2. Conclusion

The investigation into data integrity metrics involved several steps: a literature review, an ethical check, a data quality check, and an interview with a data engineer, which helped in the process of making decisions towards the final solution. To answer the how question, these steps helped create a protocol document with definitions and metrics. The ethical check showed limitations in measuring data integrity in patient data. The data quality check found inconsistencies, like dates of birth after dates of death and introduced code snippets that could run with any dataset. The interview confirmed the chosen metrics' suitability but highlighted the need for additional data for some dimensions like completeness and accuracy. Overall, the investigation achieved its goal, but certain dimensions may need more data or cross-referencing. Therefore, it can be concluded that the purpose of this sub question is achieved and can be proven with the document and results above.

4.3. Sub question 3

How can an integration into the IT landscape be incorporated, while measuring the dimension of data integrity and employing suitable visualization in a monitoring dashboard?

The purpose of this question is to research how to connect the data of AxionContinu to Power BI, and visuals that can be used for monitoring data within the prototype.

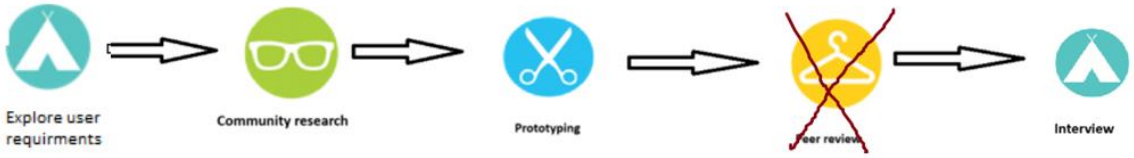


Figure 12 Methods used for the 3rd sub-question.

- Explore user requirements: New method added, to get a view of how the users will be using this dashboard.
- Community research: Researching Microsoft/ stack overflow for ways to connect the IT landscapes and methods/ visuals that helps in visualizing the errors of the prototype.
- Prototyping: Create a small PoC that would be able to visualize the errors and data quality based on the components of the automated checks.
- Peer review: This method has changed compared to the project plan, as peer review should be used to collect feedback and review the product by an external expert and not within the company.
- Interview: Conduct a short interview with the stakeholder to collect feedback and to review the work done for the purpose of improving the work, especially that they need to reuse it.

4.3.1. Results

Explore user requirements

For the explore user requirement method, a requirement analysis chapter was made and can be found in Appendix B: Research document. This document was made based on all the interviews done with all stakeholders, as in the requirements was not done based on 1 interview. An example of this is, during an interview when an interviewee says that he has a trouble with something while explaining the process, the interviewer turns that into a requirement, even though it was not mentioned as a requirement. Below, will include the top 3 requirements of each requirement as the whole table can be found in the appendix.

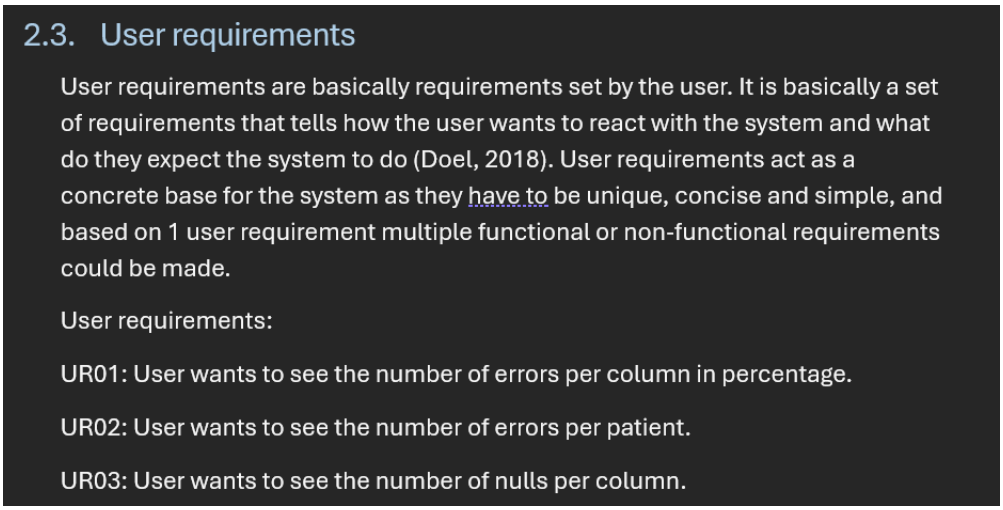


Figure 13 Screenshot of research document showing User requirements.

The figure above shows the top 3 requirements of the user requirements, where it defines what the user expects the system to do.

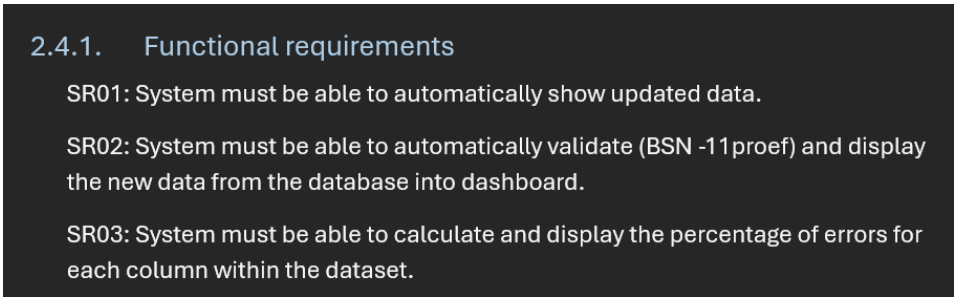


Figure 14 Screenshot of research document showing Functional requirements.

The figure above shows the top 3 functional requirements which talks about what the system should do.

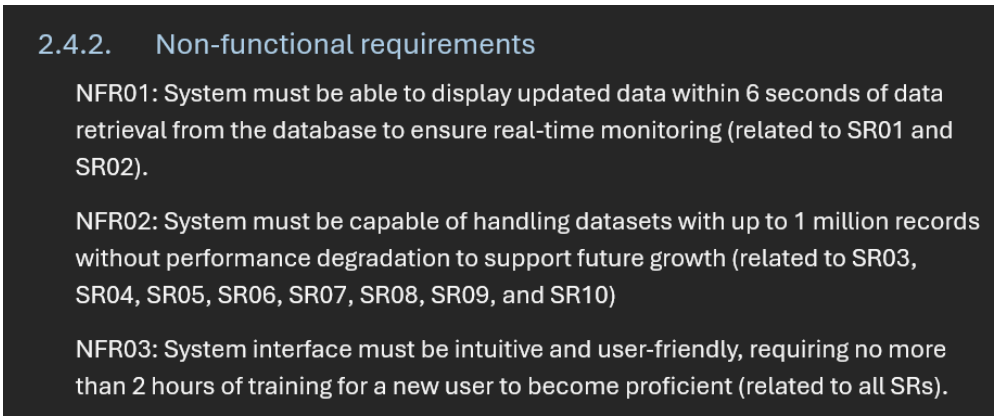


Figure 15 Screenshot of research document showing Non-functional requirements.

The figure above shows the top 3 non-functional requirements with how they relate to the functional requirements.

Community research

The first part of community research was about how to get the data into Power BI. It was discovered that Databricks has its connector within Power BI (Databricks, 2024), which can be used to get the data in, as it can be seen below.

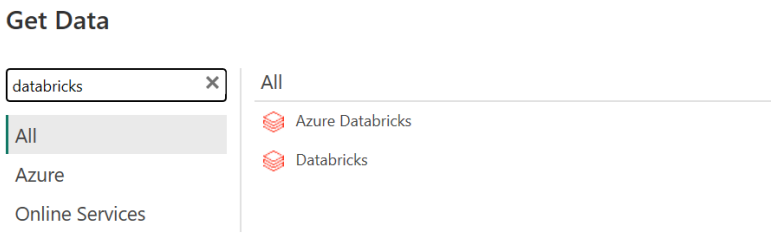


Figure 16 Showing how to connect Databricks and Power BI.

After clicking on the connector and signing in with the Username and password that was provided, navigating to the bronze warehouse and picking the tables that are needed to be loaded was done.

Additionally, before conducting community research on visualizations, Python Databricks codes were being translated into DAX Power BI codes with the assumption that they should also be visualized in Power BI, as it can be seen below.

```
DOD_Date_Statistics =
VAR MinDate = MIN('mijncaress_crsadmin_tbc'[OVL_DAT])
VAR MaxDate = MAX(mijncaress_crsadmin_tbc[OVL_DAT])
VAR NullCount = COUNTROWS(FILTER('mijncaress_crsadmin_tbc', ISBLANK('mijncaress_crsadmin_tbc'[OVL_DAT])))
RETURN
    "Min Date: " & FORMAT(MinDate, "Short Date") &
    ", Max Date: " & FORMAT(MaxDate, "Short Date") &
    ", Null Count: " & NullCount
```

Figure 17 Screenshot shows how python codes were translated into DAX codes.

Little was known about the built-in function that does the EDA (Data profiling) in Power BI. Based on the community research posted by radacad (Rad, 2019), it was discovered that in a new blank query you can profile the data and then display it in a table through a simple code, as it can be seen below.

```
= Table.Profile(mijncaress_crsadmin_tbc)
```

Figure 18 Showing profiling function in Power BI.

This code simplifies the translation of some of the coding languages, however not for the BSN 11-proef (which was hard to do in DAX) nor the date checks, as those needed to be done manually.

Prototyping

This Method will be iterated again below. For this method, a Power BI dashboard was made to visualize the error and the profiling of the data. This is one of the end products that will be delivered to the stakeholder but it's only a prototype and will not be pushed to production. However, after the community research and the data being collected, the building of the dashboard was commenced as follow:

- The data model/semantic model with couple of tables was built as it can be seen below.

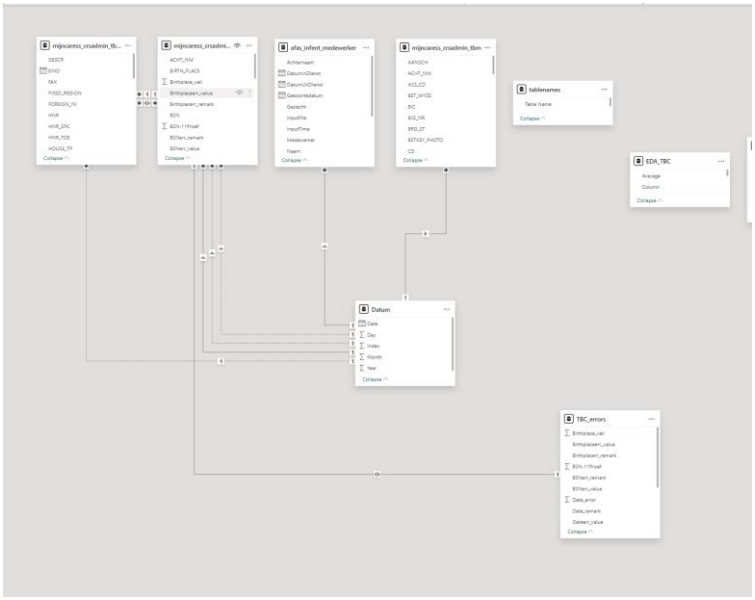


Figure 19 Showing the semantic model and relations.

- The first template of the dashboard was also made with a home page and the information or the EDA that was manually made, as it can be seen below.

Column	Min	Max	Average	StandardDeviation	Count	NullCount	DistinctCount	Uppercase	Lowercase
ACHT_NM	Aa	Ünlü			33334	0	12275	TRUE	FALSE
BIRTH_PLACE		Üzengilik			33334	777	2963	TRUE	FALSE
BSN	0	[redacted]			33334	582	32730	TRUE	FALSE
BSN-11Proef	0	1	[redacted]	[redacted]	33334	0	2	FALSE	FALSE
BSNerr_remark					33334	0		FALSE	FALSE
BSNerr_value					33334	582		FALSE	FALSE
Birthplace_vali	0	1	0,029789404	0,170008419430196	33334	0	2	FALSE	FALSE
Birthplaceerr_value					33334	777		FALSE	FALSE
Birthplacerr_remark					33334	777		FALSE	FALSE
Date_error	0	1	5,99988E-05	0,00774577304353794	33334	0	2	FALSE	FALSE
Date_remark					33334	0		FALSE	FALSE
Dateerr_value					33334	0		FALSE	FALSE

Figure 20 Shows the columns made from python.

Figure 21 Shows the home page and navigation of the dashboard.

- The BSN 11-proef was made in DAX but with some errors. As in databricks when the BSN were validated, all the BSNs were valid, and some were empty. However, in DAX, I struggled with multiplying the indexes which resulted in false results, as it can be seen below.

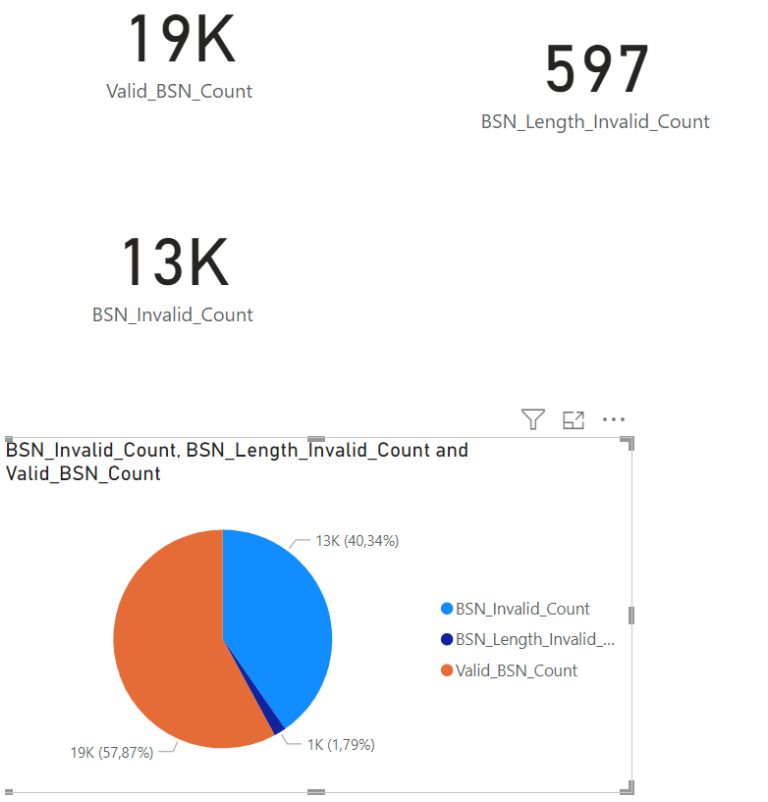


Figure 22 Shows the first visuals made for the BSN test.

As it can be seen above, it says that there are 13k BSN numbers that are not valid. However, this was fixed with the help of my mentor, as he mentioned that I need to “divide and conquer”, which got me more curious to why it’s not working. Based on a quick community and literature research, it was also discovered that there is a huge difference between the static and dynamic calculation for the 11-proef. Based on this article (Chandak, 2024), Dax measures are dynamic and are calculated on the spot based on the user interaction and are not stored. This means that it cannot handle row-by-row iterative logic very well. On the other hand, the usage of the DAX calculated columns is also not favourable, as it executes the calculation within the semantic model. This means the same as the measures, it cannot handle the 11-proef logic as it involves iterative calculations over rows and over specific ordering. However, custom columns within power query did the magic as it delivered the same results as the python notebook did. That is because the calculation happens before the data is loaded into the semantic model; this allows more complex and detailed transformations on each row. Additionally, based on this community research (Microsoft Fabric, 2018), it appears that using power query for calculations makes the performance faster on the visuals.

Interview

After building the prototype, an interview was mode with the mentor/ boss Maikel, to validate the prototype and collect feedback for iteration. During the review, the dashboard was presented, but the boss expressed dissatisfaction, stating that it primarily showcased the exploratory data analysis (EDA) or data profiling rather than highlighting errors. Additionally, it was noted that I continued to focus on the analysis phase. I explained to the boss that Fontys' preference is to continuously iterate between analysis and prototyping. For instance, they don't share the belief that the analysis phase can be considered complete, as desired by the stakeholder. Nonetheless, feedback was also collected on how they would like the errors to be visualized with the requirements of the drill through, as it can be seen below.

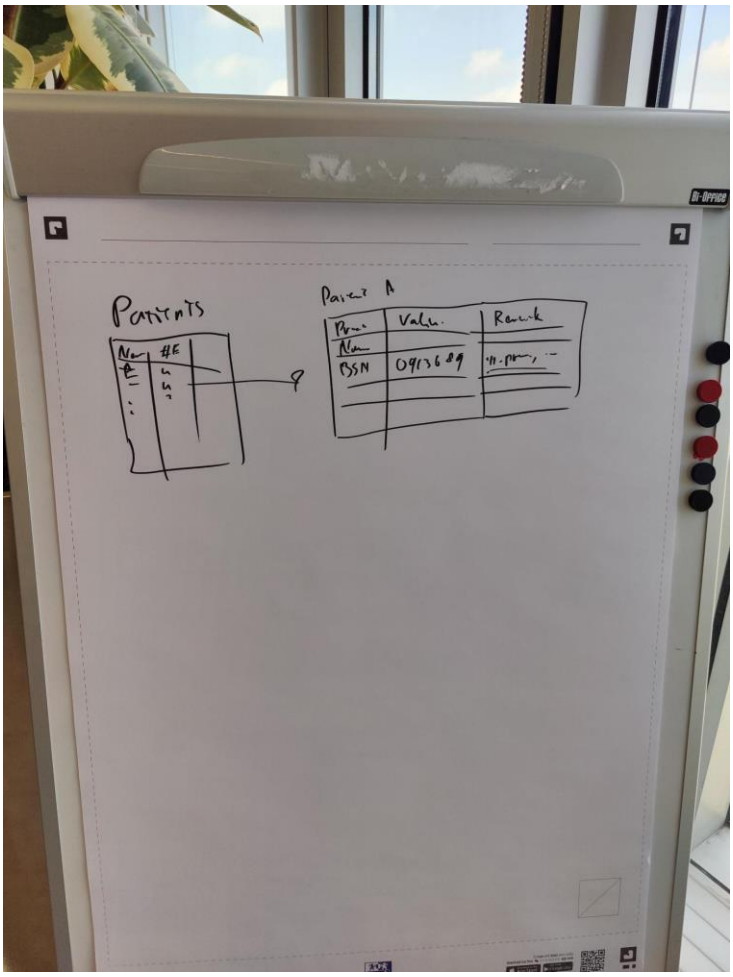


Figure 23 Shows the wishes of the stakeholder.

The figure above shows the requirement of the stakeholder which can be summarized as follow:

1. First page displays the patients and the number of errors per row (includes BSN check, Date Check, Ucase checks and etc).
2. From the first page a drill through to the second page which shows the errors and their values in this format:

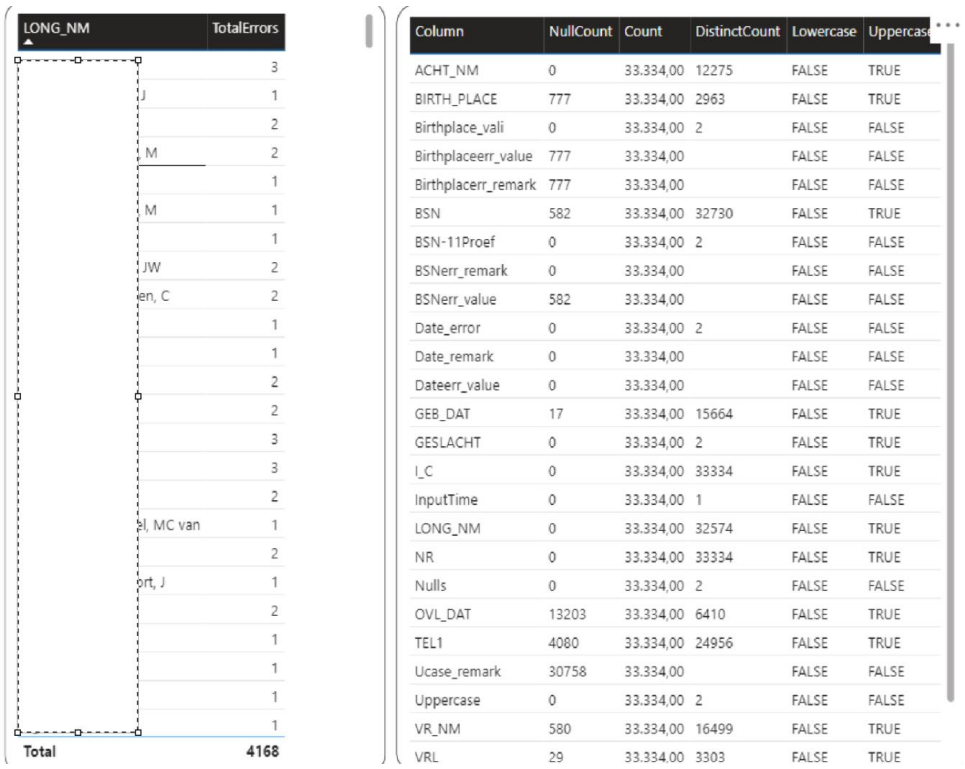
Table 1 Shows the wished pivoted format.

BSN Error	1231232134	BSN does not pass the 11 proef
Ucase error	hein-jan	The name didn't start with a capital letter.

3. Another visual that displays the error and the number of patients that has that error and when you drill through you see all the patients who has that error in their record.

Prototype part 2

Based on the review and feedback collected, the dashboard has improved to the wishes of the stakeholder, as it can be seen in the figure below.



LONG_NM	TotalErrors
	3
J	1
	2
M	2
	1
M	1
	1
JW	2
en, C	2
	1
	1
	2
	2
	3
	3
	2
el, MC van	1
	2
ort, J	1
	2
	1
	1
	1
	1
Total	4168

Column	NullCount	Count	DistinctCount	Lowercase	Uppercas
ACHT_NM	0	33.334,00	12275	FALSE	TRUE
BIRTH_PLACE	777	33.334,00	2963	FALSE	TRUE
Birthplace_vali	0	33.334,00	2	FALSE	FALSE
Birthplaceerr_value	777	33.334,00		FALSE	FALSE
Birthplacerr_remark	777	33.334,00		FALSE	FALSE
BSN	582	33.334,00	32730	FALSE	TRUE
BSN-11Proef	0	33.334,00	2	FALSE	FALSE
BSNerr_remark	0	33.334,00		FALSE	FALSE
BSNerr_value	582	33.334,00		FALSE	FALSE
Date_error	0	33.334,00	2	FALSE	FALSE
Date_remark	0	33.334,00		FALSE	FALSE
Dateerr_value	0	33.334,00		FALSE	FALSE
GEB_DAT	17	33.334,00	15664	FALSE	TRUE
GESLACHT	0	33.334,00	2	FALSE	TRUE
I_C	0	33.334,00	33334	FALSE	TRUE
InputTime	0	33.334,00	1	FALSE	FALSE
LONG_NM	0	33.334,00	32574	FALSE	TRUE
NR	0	33.334,00	33334	FALSE	TRUE
Nulls	0	33.334,00	2	FALSE	FALSE
OVL_DAT	13203	33.334,00	6410	FALSE	TRUE
TEL1	4080	33.334,00	24956	FALSE	TRUE
Ucase_remark	30758	33.334,00		FALSE	FALSE
Uppercase	0	33.334,00	2	FALSE	FALSE
VR_NM	580	33.334,00	16499	FALSE	TRUE
VRL	29	33.334,00	3303	FALSE	TRUE

Figure 24 Shows the new iteration of the wished visual.

In the figure above, on the left side table, the patients are displayed with the count of total errors per patient and once you drill through a patient it takes you to a different page as required. This was achieved by following the basic programming mindset, as follow:

1. Create custom columns that captures what is needed and display it as 0 or 1(i.e., Incase BSN failed then make it 1 and if it didn't then 0).
2. A measure or a calculated column can be made for this, where it checks if the BSN is 1 then adds a remark (i.e., BSN failed 11-proef).
3. Final step is to create a measure that sums all the error together and call it "Total Errors" which then can be seen in the figure above on the left side.

LONG_NM	[redacted]
[redacted]	
Birthplaceerr_value	?
Birthplacerr_remark	Can contain the following : @, *, #, ?, !, &, `, |, \, /, (, „, $, _
BSNerr_value	
BSNerr_remark	Invalid length or failed 11 proef
Dateerr_value	DOD-No error
Date_remark	DOD-No remark
First Ucase_remark	Check Ucase of the following columns: 1. Long_NM, 2. Geslacht, 3. Voornaam en achternaam 4. Birthplace

Figure 25 First iteration of the drillthrough function based on the requirements.

As it can be seen above, this row or this patient has BSN missing and has a "?" as his birthplace. Additionally, also based on the review, the third requirement is also made as it can be seen below.

BSN-11Proef	Count of LONG_NM
0	32737
1	597

Birthplace_vali	Count of LONG_NM
0	32341
1	993

Date_error	Count of LONG_NM
0	33332
1	2

Nulls	Count of LONG_NM
0	2541
1	30793

Figure 26 Showing the number of errors in separate visuals.

This visual is showing the count of patients per error. Furthermore, during the building of the dashboard, bookmarks were discovered and were used for the design of the dashboard. The figure below, shows the filters on the left side of the screen, these filter uses a bookmark to save space.

Figure 27 Showing the Filter button.

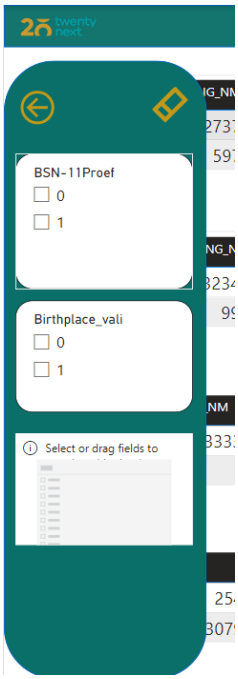


Figure 28 Showing the power of bookmarks to save space on the dashboard.

Additionally, a new feature was discovered while building the dashboard and that is showing the applied filter or filters in use without taking much space in the dashboard page, as it can be seen below.

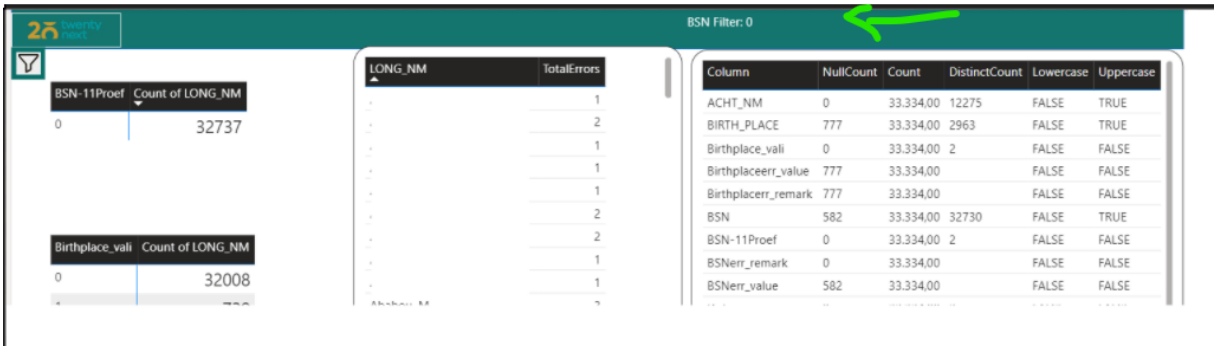


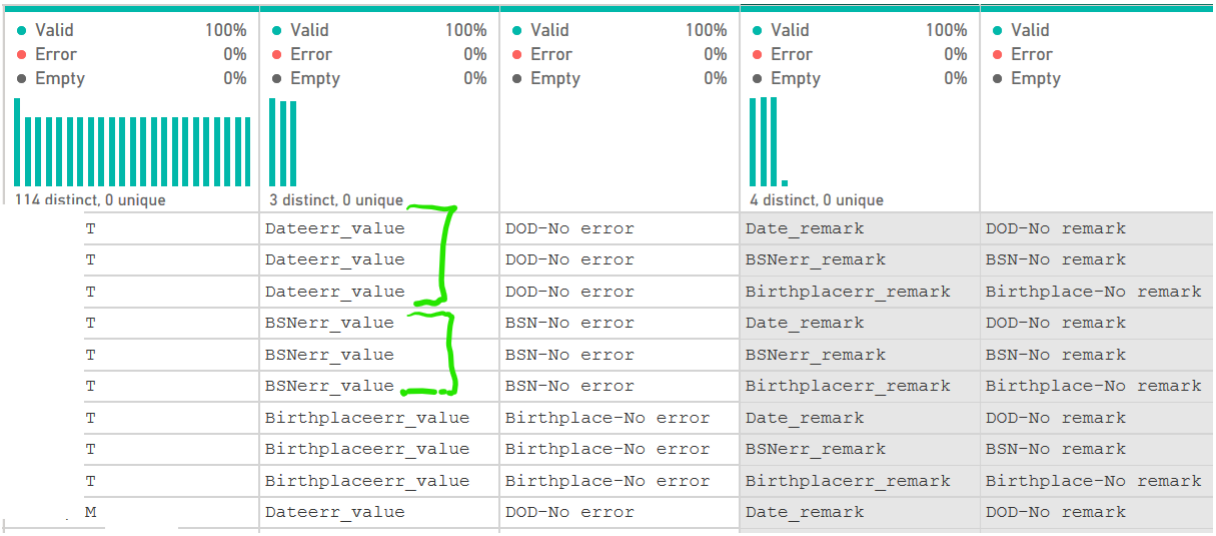
Figure 29 Showing new ways to display filtered values.

These filters remain empty when no filtering is applied but become visible when any filtering is activated. This Dashboard will go through a third iteration.

Prototype part 3

During the third iteration, feedback was gathered through one-on-one sessions with the company's mentor. We discussed the possibility of restructuring the table into the required format, as mentioned in the interview section above, addressing the

challenges encountered in doing so. The challenge that was faced is the duplication of the rows when pivoting (Chereden., 2021) the table, as it can be seen below.

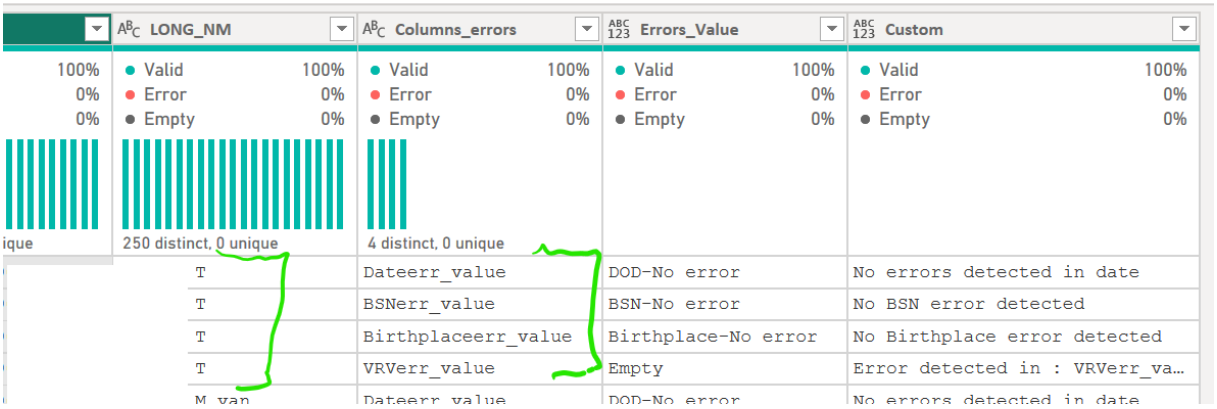


Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid Error Empty
114 distinct, 0 unique	3 distinct, 0 unique		4 distinct, 0 unique	
T	Dateerr_value	DOD-No error	Date_remark	DOD-No remark
T	Dateerr_value	DOD-No error	BSNerr_remark	BSN-No remark
T	Dateerr_value	DOD-No error	Birthplacerr_remark	Birthplace-No remark
T	BSNerr_value	BSN-No error	Date_remark	DOD-No remark
T	BSNerr_value	BSN-No error	BSNerr_remark	BSN-No remark
T	BSNerr_value	BSN-No error	Birthplacerr_remark	Birthplace-No remark
T	Birthplaceerr_value	Birthplace-No error	Date_remark	DOD-No remark
T	Birthplaceerr_value	Birthplace-No error	BSNerr_remark	BSN-No remark
T	Birthplaceerr_value	Birthplace-No error	Birthplacerr_remark	Birthplace-No remark
M	Dateerr_value	DOD-No error	Date_remark	DOD-No remark

Figure 30 Showing the struggles during the pivoting.

Per error, each error was duplicated three times per patient. To solve this problem, the remark columns were deleted and later added into a single custom column, eliminating the need for pivoting, as shown below.

```
= Table.AddColumn(#"Renamed Columns1", "Custom", each if [Errors_Value] = "DOD-No error" then "No errors detected in da
else if [Errors_Value] = "BSN-No error" then "No BSN error detected"
else if [Errors_Value]= "Birthplace-No error" then "No Birthplace error detected"
else if [Errors_Value]= "VRV-No error" then "No Voorvoegsels error detected"
else "Error detected in : " & [Columns_errors])
```



	LONG_NM	Columns_errors	Errors_Value	Custom
100% 0% 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%
ique	250 distinct, 0 unique	4 distinct, 0 unique		
	T	Dateerr_value	DOD-No error	No errors detected in date
	T	BSNerr_value	BSN-No error	No BSN error detected
	T	Birthplaceerr_value	Birthplace-No error	No Birthplace error detected
	T	VRVerr_value	Empty	Error detected in : VRVerr_va…
	M van	Dateerr value	DOD-No error	No errors detected in date

Figure 31 Showing the wanted results without the need of pivoting.

Moreover, a drawback was identified in Power BI where null values are not displayed. Consequently, to remedy this issue and ensure the error is visible in the table, null values had to be converted into a string labeled "Empty". Otherwise, the table would

have shown only 3 rows instead of 4, as demonstrated below.

LONG_NM	Columns_errors	Errors_Value	Custom
Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%	Valid 100% Error 0% Empty 0%
250 distinct, 0 unique	4 distinct, 0 unique		
T	Dateerr_value	DOD-No error	No errors detected in date
T	BSNerr_value	BSN-No error	No BSN error detected
T	Birthplaceerr_value	Birthplace-No error	No Birthplace error detected
T	VRVerr_value	Empty	Error detected in : VRVerr_va…
M van	Dateerr value	DOD-No error	No errors detected in date

Figure 32 Showing that empty rows had to include a string "Empty" for the error to show.

After the transformation of the data was done, it was time to improve on the design of the dashboard. The design has changed compared to figure ***. As it can be seen below, the new design includes percentages of errors per error on the left side of the screen. The middle section includes the amount of errors per record and the colors varies from green (0 errors) to red (>4).

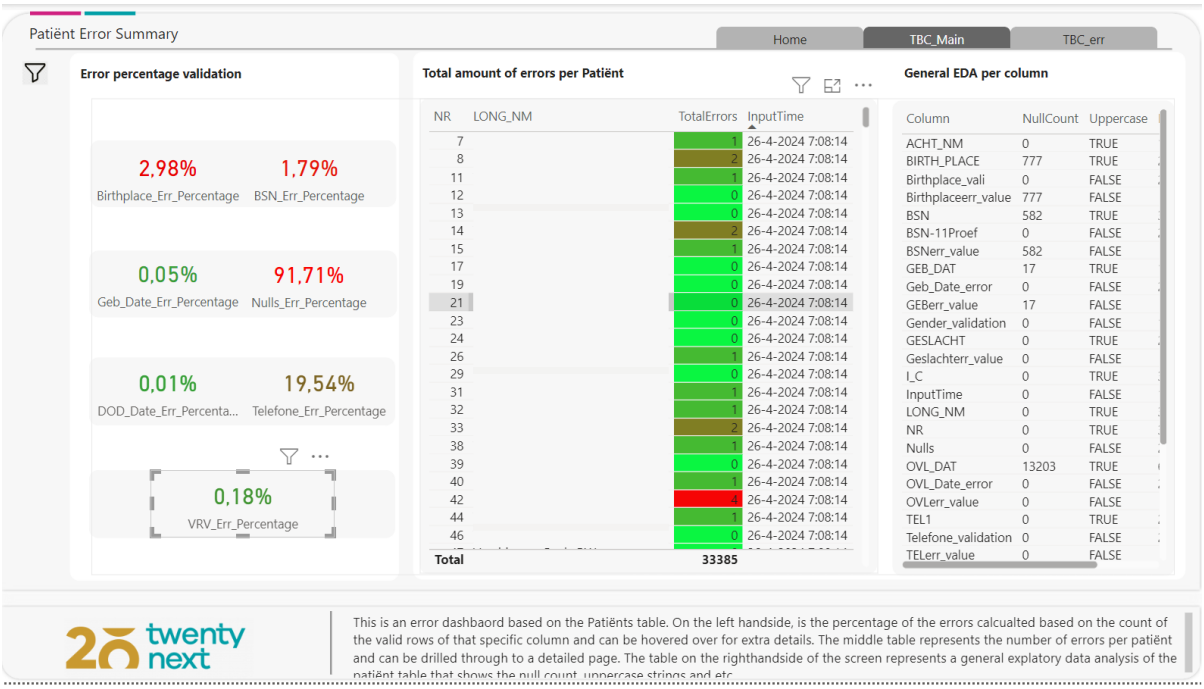


Figure 33 Showing the main page of the final iteration of the dashboard.

Once you drill through the record to check the exact error, a page will appear with the pivoted table that includes the details based on the wished format.

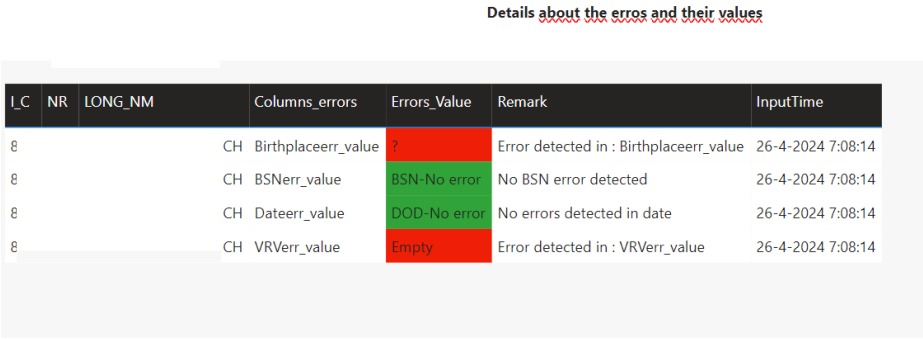


Figure 34 Showing the drill through table format.

As it can be seen above, this record contains errors like the "?" in the birthplace. These errors are colored in red, so that it automatically captures the attention of the user. On further iteration, the same calculations were done on the addresses table and were added to the visuals as it can be seen below.

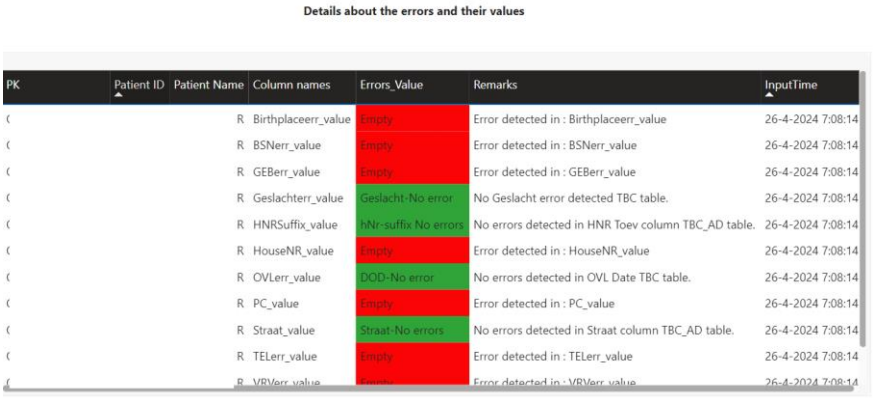


Figure 35 Showing the final iteration of the drill through table.

Another iteration was to fix the number of tables in the dashboard after adding the address table. It was decided to merge these 2 tables into one table that is called "Main_errors_table", as it can be seen below.

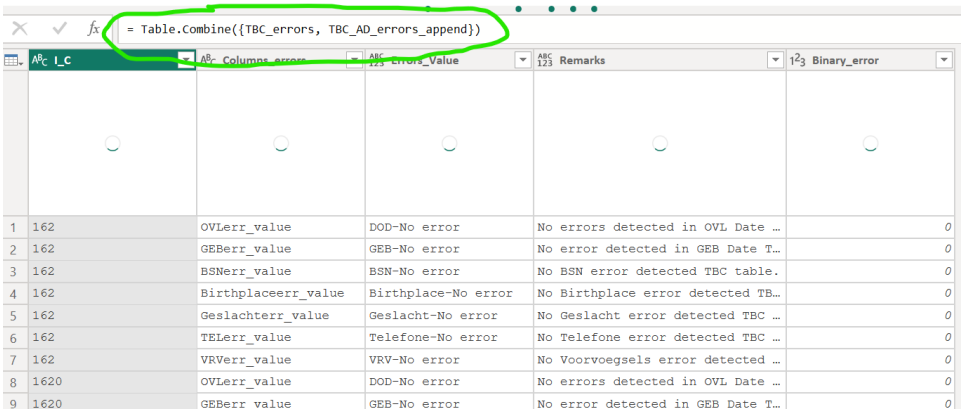


Figure 36 Showing the merge of the tables.

And then we can disable the load of these tables but not the refresh so that we get new data, by clicking on the tick, as it can be seen below.

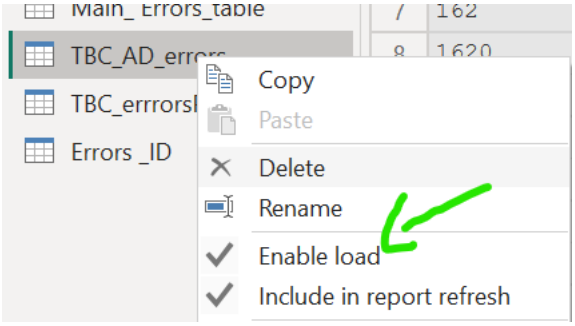


Figure 37 Showing how to disable dataload.

4.3.2. Conclusion

This sub question was quite a heavy question due to its complexity of combining multiple aspects: IT integration, requirements collection, data integrity measurements and visualization in a monitoring dashboard. However, the research presents a thorough examination of data integration, visualization, and prototype development in Power BI. Through iterative prototyping (that contributes to the quality of the work), requirements analysis and stakeholder feedback, the dashboard evolved to effectively address stakeholder requirements, particularly in error visualization and data profiling. Challenges such as data restructuring and null value handling were overcome with innovative solutions (out of the box thinking). Therefore, it can be concluded that the purpose of this sub question is achieved and can be proven with the document and results above.

5. Conclusion and recommendations

In this chapter, the conclusion and recommendations of this project are stated.

5.1. Conclusion

The project began with identifying several issues detailed in the problem analysis chapter. These issues were consolidated into a central research question: "How can a methodical instrument be designed to measure data integrity effectively, incorporating flexible parameters aligned with business rules, and ensuring consistency and reliability in the assessment process?" This question served as the foundation for the research and guided the investigative focus throughout the project. From this main question, three sub-questions were formulated and investigated to address the overarching inquiry. The initial sub-question focused on identifying various frameworks and establishing a protocol for the ETL process conducted by a data engineer. The second sub-question involved identifying metrics to augment the protocol, analyzing existing data for errors and patterns, and converting these patterns into rules (such as the BSN 11-test or character count for postcodes or BSNs). The final sub-question centered on the development of the dashboard based on the requirement analysis. Derived from the outcomes of these sub-questions, Twentynext has acquired a monitoring tool prototype, offering insights into the data integrity of AxionContinu. Furthermore, a notebook code snippet has been developed to analyze various datasets, demonstrating adaptability to new or updated data. Consequently, it is evident that the primary research question has been successfully addressed and resolved. Nevertheless, the thesis remained largely aligned with the project plan, adhering to the outlined deliverables and the criteria specified in the MoSCoW tables. The 2 sole deviations occurred in the approach to addressing the third sub-question. Initially, the "Peer review" method was proposed, but it became apparent that this method was intended for external expert validation rather than internal organizational use and due to confidentiality agreements, further utilization of this method was excluded, necessitating a shift to the "Interview" method instead. The 2nd deviation was the requirement analysis method, as it was added as a future orientation, to ensure how the dashboard is expected to be used. Moreover, concerning the overall completion status of this internship project, it stands at approximately 80%. All essential tasks outlined in the MoSCoW table have been successfully completed within the designated timeframe. However, certain "Should" tasks were not feasible due to time constraints but will be recommended for future consideration. In contrast, regarding the project's completeness from Twentynext's perspective, only about 20% has been addressed, leaving a significant portion yet to be accomplished. This project was initially scheduled to commence within approximately six months or by the end of 2024 by a team that consist of more than one employee, and thus, this internship's scope was limited to addressing the consistency aspect of data quality dimensions.

5.2. Recommendations

Based on this project and the specific scope of it, it is recommended for the next steps to choose a framework for data governance, since that data quality is just a small fraction of it and data governance covers more, including security and compliances.

Secondly, it is recommended to utilize the monitoring functions within databricks, as it offers more value regarding dealing with ETL processes and visualizing them, which will reduce the need of Power BI and increases the time of real time data monitoring. As it can go (Source-> bronze layer -> monitor data) within the same environment and not (Source -> Bronze layer-> external Power BI environment). Additionally, databricks has some metrics that helped in the development of the advised protocol, which can also be improved further.

Thirdly, for future heavy row level calculations, it is highly recommended to use power query or advanced editor, as it increases the performance and visualizes correct results.

Fourthly, it is recommended to continuously develop and update the protocol based on feedback from the engineers and new industry standards, to maintain the relevance and effectiveness of the protocol.

Finally, expanding the scope of the data integrity project, as not just looking into the other tables for consistency but following the MoSCow and looking at other aspects such as validity and completeness, etc.

Nonetheless, the advice is based on the research and the scope of this project, so ultimately the project can be developed further, dashboard can be used for testing, feedback can be collected and processed by the roll of the data analyst.

Assignment evaluation

In this chapter the assignment including the goal, strategies and methods will be evaluated below.

Problem and goal

The project started with the problem of that an employee of AxionContinu not being able to get consistent and valid patient's data. This was due to human input errors (i.e., switching values of DOB and DOD) and not having a tool that could flag these errors. These issues were approached with a goal of "having an end product that finds patterns within the data, validates and flags the data errors based on these patterns and gives insight into these errors."

Therefore, to solve these issues, a proof-of-concept dashboard, Jupiter notebook and protocol were made. The protocol provides technical information to the data engineer on metrics to use to validate the data during the ETL process within data bricks. The Jupyter notebook is collecting the data and running it through the same code chunk to find patterns, errors and validate the data (i.e., 11-proef test). The dashboard is a code translation of the notebook (excluding the EDA) and it visualizes the errors per columns per patient. Based on these end products, Twentynext now has a tool that can visualize the errors and support the data quality enhancement of AxionContinu.

Research strategies and methods

During the research part of the project, all the ICT research strategies were used based on the dot framework, at least once per research question. Each research question had at least 3 different research strategies used. From these strategies, multiple methods were used, such as Community research, literature study, interview, prototype, etc.

Nonetheless, there was a bit of a struggle in finding a suitable method to use in the last research question, especially during the beginning of the project. Later on, it was brought to attention that the "Peer review" method was not fitting with the question, therefore it was changed and elaborated on the report on why it has changed. Additionally, the same goes for the "requirements analysis method" as it was added and elaborated on why it was added.

Guidance moments

Due to the complexity and size of this project, guidance was one of the mandatory things that the intern needed the most. The first guidance moment was when the scope of the project was big and the intern had to reduce it to fit the timeline of 18 weeks, consequently, the MoSCow tables in the project plan were developed to limit that and that was given by my school mentor Rink Lycklama. The second guidance moment was during the first research time. Where the intern couldn't see the bigger picture of the data quality and the difference between user data quality and data file/folder (CSV format or excel) data quality. It was a struggle, as the intern only knew about the concept of data quality during the 4th semester (AI specialization) and was trying to use that knowledge in this project. Therefore, guidance was needed and delivered by the company mentor Maikel Maasakkers and as a result of that patterns were found and documented. The third guidance moment that I needed was with the report, as the intern knows that his reporting skill needs improvements especially when it comes to summarizing. Therefore, guidance was given by my first assessor Rink Lycklama and further consistent guidance on this matter was given to me by a co-worker called Ni Kang. Based on the feedback, it was significantly noticed that the document is becoming more concrete and professional. The fourth guidance moment that was needed was also with writing as the intern couldn't put the description into concrete words. Guidance was given by Rink Lycklama that the problem needs to be visualized and drawn on paper then translated to words. This feedback was implemented and helped as the intern does the visualization trick with codes and programing languages and not with writing, so that was a learning curve. The fifth guidance moment was when the intern was translating python codes into DAX and was struggling with the 11-proef test. Guidance was given by the company's mentor which made the intern more curious about the source of the problem. Based on that guidance, the intern found the source and the reason for it not working and made it work. The final guidance moment was when the intern couldn't pivot the tables to the requirements of the stakeholder. Guidance was delivered by another co-worker called Huck Thönissen who was specialized with Power BI within Twentynext. The intern did not collect feedback from Huck but talked to him out loud about the problem and then figured it out himself, as sometimes you need to talk out load to think and hear yourself better, and then the tables were pivoted as wished.

Learning experience

During this project I have learned a lot, and it can be summarized into the following:

- Usually within Fontys we work in groups but with the complexity and size of this project and having to work independently, I learned to work independently in such a huge project.
- I increased more and improved on my knowledge about data quality that I learned about during my AI specialization.
- I have learned how to write a report and stick to the point by visualizing it.
- I have learned more about the Dot framework and how to pick suitable strategies and methods.
- I have widened my knowledge within databricks and the architecture structure that accompanies.
- I have also learned about data standards and tests like the BSN 11 proef test as that was translated from a mathematical formula into a python and a DAX code.
- I have also learned how to make a generic piece of code that continue to work based on multiple tables and not tailored to a specific table.
- I have learned about new features within Power BI.
- I have also learned about the calculation limitations of Power BI.
- I have improved my time management skills.
- I have learned how to approach and solve conflicts when they rise within an organization.
- I have learned to ask directly and that “you won’t get anything unless you ask” concept.
- Most important lesson that I have learned is to make plan B, C, D and E and not just wait for A, as that is time consuming and its always better to collect feedback from multiple sources.

Reflecting on the beginning of the internship and considering the points mentioned above, I have demonstrated my growth not only to myself but also to my school assessor, work assessor, and colleagues. This showcases my enthusiasm and proactiveness towards the project, my ability to process feedback, and my commitment to retaining these lessons for life.

Personal reflection

Reflection based on the assessment form

In the assessment form, there are 6 assessment dimensions that I will be assessed on, and I will be using below to reflect on myself. The grading goes as follow:

- Undefinded (U)
- Orienting (O)
- Beginning (B)
- Proficient (P)
- Advanced (A)

On each assessment dimension, I will elaborate what I have done, what went well and what could be done better with a grading that I will give myself.

Professional Duties

Within this section, I would grade myself a “Proficient” here. This is because I have carried out the whole assignment professionally and on a high-level touching all the phases of the IT life cycle, as follows:

- Analysis: where I analyzed the core of the problem and the data itself can be seen in chapter 3 and research question 1.
- Design: where I designed what is needed to be an end product and how and etc and can be seen in the project plan and research document.
- Realize: where I realized the protocol and the notebook can be seen in the notebook itself and protocol in research question 2.
- Advice: where I gave advice on what the next steps can be and that can be seen in the advice/ research document.
- Manage&control: where I made a monitoring dashboard and described roles and responsibilities within the advised protocol document.

I am not praising myself here but everyone I worked with or for even my mentors from previous semesters always tell me that I always deliver a very thorough research and reporting. I know that I do that, but I don’t feel that I do that as there is always something to learn and to do. Additionally, within my professional duties, I also kept in mind that not all information can be shared and with every step I did while writing the thesis, I kept the NDA in mind.

Situation-Orientation

Within this section, I would grade myself a “Advanced” here. Not that I just can work in a complex technical environment, but also, I can use previous knowledge. As I have been strictly developing myself within the data engineering world and I have learned about data science in my 4th semester and my first internship was a data analysis project, I have used all these 3 in this project. Being a data engineer, helped in developing the dashboard and the protocol. The data science, helped in developing the metrics, validation, and the whole data quality theory. The data analysis part helped in finding the patterns and errors associated with it. Additionally, are and will be used, as the protocol is already shared with the data engineer and the dashboard will be developed further into production level, as it is only a proof of concept for now.

Future-Oriented Organization

Within this section, I would grade myself a “Proficient” here. This is because of the fact that the last submission of the project plan was perfect. As it covered the whole planning process, monitored the projects execution, documented the approaches and strategies perfectly, mentioned the risks and the quality of the end products/ what is expected as an end product. In addition to that, within the recommendation and advice, I gave a financial advice about the usage of databricks instead of Power BI, I also thought about a future performance proof and sustainable way to write codes on Power BI and advised it. Furthermore, the Python code that was made is also future oriented, as when making it I made sure that it will not crash no matter what type of table you run through it.

Investigative problem solving

Within this section, I would grade myself a “Advanced” here. As since the beginning of the project, I identified the problems, made a requirement analysis chapter, defined a scope which was reduced later, used effective DOT framework methods to approach the research questions, used the results of the research to create valuable end products such as jupyter notebook and etc.

Additionally, whenever a problem was emerged (within Power BI, 11 proef, or notebook) I was able to tackle it methodologically and logically. A perfect example of this is having the programing mind set during the realization of the Power BI as the values were changed to zeros and ones, when this challenge was recognized.

Personal leadership

Within this section, I would grade myself a “Advanced” here. As I have already been working in my wanted field of data engineering for the last 2 years and that is what helped a lot also in this project. Being already in my wanted position doesn’t mean that there is nothing to learn, but it means that there is so much to learn like spark language or other tools within azure like Azure AD and etc. I have been also very entrepreneurial during this project, and I think that is already proved and spoken by itself, by my documents, work, and learning experience that is mentioned above is the perfect proof for this. I always seek for improvements and learning opportunities, and this is one of my main requirements for either my internship company or my other job, for a after Fontys. As we can never learn enough!

Targeted interaction

Within this section, I would grade myself a “Advanced” here. This is due to a lesson I’ve learned of being direct, to the point, and how valuable my time is. I always consult the project members to get any information I need and plan meetings with them when I need their support. I participated and shared everything I do every day during the standup meetings. I communicated clearly, vividly, and professionally with my company’s mentor when a conflict raised and achieved the desired impact. I bothered my school mentor so much with my teams’ messages to collect feedback. When I noticed that that will be late, I automatically reached out to my co-workers to help give feedback on the documents and to improve it. Additionally, I collaborated closely with my company’s mentor during the documentation of the patterns as I was struggling with seeing the bigger picture and through our communication we got the desired outcome. Lastly, the fact that I found this project and internship company and got accepted, 1 week before the deadline, is enough of a big proof from this learning outcome.

Overall performance

In general, this was one of the hardest semesters I have ever experienced. The only 2 hard challenges I faced this semester were the project as it is super complex and challenging for just one person to work on it (including the stakeholder’s requirements) and the second challenge was me to myself. As I doubted every work I did and double checked it and added extra methods because it just felt not concrete enough and continued iterating my work, resubmitting on canvas, and always updating my feedpulse. It is true that I have learned a lot in this project, all stakeholders are happy with the results, and I have completed the scope within the limited time and showed my professional work. In the end, the delivered end products are what I take pride in, as they were successful. As I often emphasize in every reflection or learning outcome document, while I may have graded myself as 'Advanced ++' or 'Outstanding++', this merely reflects the current work completed and does not imply there's nothing more to learn or that I know everything about a topic. Unlike the others, as a teacher once told me, 'I have a bachelor’s degree, I am done learning' to which I responded, 'learning never stops' and that’s what I believe in.

Bibliografie

Linkedin. (2024, 02 19). *How can you design data quality test cases?* Retrieved from Linkedin: https://www.linkedin.com/advice/0/how-can-you-design-data-quality-test-cases-skills-data-quality

Agile alliance. (2022, 05 26). *What is Agile Software Development?* Retrieved from Agile alllinace: https://www.agilealliance.org/agile101/

AI, C. (2022, 05 2). *Importance Of Data Quality In The Age of AI*. Retrieved from Medium: https://cetasai.medium.com/importance-of-data-quality-in-the-age-of-ai-f91ad2c49c7d

Alexsoft. (2019, 10 17). *Data Quality Management: Roles, processes, tools.* . Retrieved from AltexSoft: https://www.altexsoft.com/blog/data-quality-management-and-tools/

Aliaga, A. (2023, 12 20). *introduction to Databricks Lakehouse monitoring - Antonio Aliaga - Medium.* Retrieved from Medium. : https://medium.com/@antaliagacortes/introduction-to-databricks-lakehouse-monitoring-aebeddf013b5

Anwar, M. (2024, 4 4). *5 Crucial Best practices for ensuring data quality in healthcare. Astera.* . Retrieved from Astera: https://www.astera.com/type/blog/managing-data-quality-in-healthcare/

AxionContinu. (2023, 12 4). *Over ons.* . Retrieved from AxionContinu: https://www.AxionContinu.nl/over-ons#actueel

Aziz, F. (2023, 07 11). *Requirement analysis process in software development.* Retrieved from InvoZone: https://invozone.com/blog/importance-of-requirement-analysis-process-in-software-development/

Benson, P. (2008, 07 16). *ISO 8000 the International Standard for Data Quality* . Retrieved from The MIT 2008 Information Quality Industry Symposium: http://mitiq.mit.edu/IQIS/Documents/CDOIQS_200877/Papers/13_01_5A-1.pdf

Benson, P. (2020, 02 29). *ISO 8000: A new International Standard for Data quality, by Peter Benson* . Retrieved from Data Quality Pro. Data Quality Pro. : https://www.dataqualitypro.com/blog/iso-8000-new-international-standard-data-quality

Bonestroo, W. M. (2018). *ICT Research Methods.* Retrieved from HBO-i, Amsterdam. ISBN/EAN: 9990002067426.: https://ictresearchmethods.nl/about/

brook, C. (2023, 05 08). *What is a Health Information System?* Retrieved from Digital Guardian. : https://www.digitalguardian.com/blog/what-health-information-system

Business insights blog. (2021, 02 4). *What is data integrity and why does it matter?* Retrieved from Harvard business school: https://online.hbs.edu/blog/post/what-is-data-integrity

cambridge. (2024). *beacon*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/beacon

cambridge. (2024). *bolstering*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/bolster?q=bolstering

cambridge. (2024). *consolidating*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/consolidate?q=consolidating

Cambridge. (2024). *decipher*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/decipher

cambridge. (2024). *harness*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/harness

cambridge. (2024). *hinder*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/hinder?q=hindering+

cambridge. (2024). *leverage*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/leverage

cambridge. (2024). *multifaceted*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/multifaceted

cambridge. (2024). *robust*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/robust

Chandak, A. (2024, 03 26). *Understanding Visual Calculations in Power BI: Revolutionizing Data Analysis.* . Retrieved from Medium.: https://medium.com/microsoft-power-bi/understanding-visual-calculations-in-power-bi-revolutionizing-data-analysis-1c15c943243e

Chereden., D. (2021, 12 15). *How to create a “Pivot Table” in Power BI | Power BI Basics | Easy for Excel Users [Video].* . . Retrieved from YouTube: https://www.youtube.com/watch?v=lohe7HD98wk

Cox, A. (2017, 09 06). *Business Requirements vs Functional Requirements? Who Cares?* Retrieved from EN. Netmind. : https://netmind.net/en/business-vs-functional-requirements-who-cares-en/

Cox, A. (2024, 05 13). *Software Requirements Specification (SRS):* . Retrieved from definition, example, how to write, & more. : https://www.inflectra.com/Ideas/Topic/Requirements-Definition.aspx

Databricks. (2024, 3 3). *Connect Power BI to databricks.* Retrieved from Databricks on AWS.: https://docs.databricks.com/en/partners/bi/power-bi.html

Databricks. (2024, 04 10). *Databricks*. Retrieved from What is a Medallion Architecture? : https://www.databricks.com/glossary/medallion-architecture

Databricks. (2024, 3 4). *Monitor metric tables.* Retrieved from Databricks on AWS. : https://docs.databricks.com/en/lakehouse-monitoring/monitor-output.html

Democracy westlancs gov uk. (2012, august). *DATA QUALITY PROTOCOL*. Retrieved from https://democracy.westlancs.gov.uk/Data/Audit%20&%20Governance%20Committee/201209251900/Agenda/023972_DQPROTOCOL.pdf

Doel, R. V. (2018, 01 12). *User Requirements*. Retrieved from https://perfval.com/user-requirements/

Foot, C. (2022, 5 6). *8 proactive steps to improve data quality.* . Retrieved from Data Management. : https://www.techtarget.com/searchdatamanagement/feature/Proactive-practices-for-data-quality-improvement

Ganjhu, P. K. (2023, 05 24). *The DAMA-DMBOK Functional Framework: A Comprehensive approach to effective data management.* Retrieved from Medium: https://pawankg.medium.com/the-dama-dmbok-functional-framework-a-comprehensive-approach-to-effective-data-management-3de06af6

GfG. (2023, 5 17). *Difference between Data Administrator (DA) and Database Administrator (DBA).* . Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/difference-between-data-administrator-da-and-database-administrator-dba/

HU, K. (2023, 05 10). *How to set up data quality Tests*. Retrieved from Metaplane: https://www.metaplane.dev/blog/how-to-set-up-data-quality-tests

IBM documentation. . (2021, 03 05). *Data quality methodology*. Retrieved from IBM: https://www.ibm.com/docs/en/iis/9.1?topic=practices-data-quality-methodology

ICTinformatiecentrum. (2023, 11 08). *Data governance framework*. Retrieved from ICTinformatiecentrum: https://www.ictinformatiecentrum.nl/data-management/data-governance/data-governance-framework

Incept Data Solutions, Inc. . (2023, 02 17). *WHICH DATA GOVERNANCE FRAMEWORK SHOULD YOU CHOOSE?* . Retrieved from LinkedIN: https://www.linkedin.com/pulse/which-data-governance-framework-should-you-

ISO 8000. (2022, 02). *Data quality*. Retrieved from Part 2: Vocabulary. : https://www.iso.org/obp/ui/#iso:std:iso:8000:-2:ed-5:v1:en

ISO/TS 8000. (2012, 11 3). *Data quality*. Retrieved from Part 311: Guidance for the application of product data quality for shape (PDQ-S).: https://www.iso.org/obp/ui/#iso:std:iso:ts:8000:-311:ed-1:v1:en

Karl. (2024, 2 9). *Audit Analytics: types, benefits and use cases.* . Retrieved from Caseware Canada. : https://www.caseware.com/ca/resources/blog/audit-analytics-types-benefits-and-use-cases/

LakeFS. (2024, 03 11). *What is Data Quality?* Retrieved from Definition, Framework & Best Practices: https://lakefs.io/data-quality/

LakeFS. (2024, 3 11). *Data Quality Monitoring: key Metrics, techniques & Benefits.* . Retrieved from Git For Data - lakeFS.: https://lakefs.io/data-quality/data-quality-monitoring/

LakeFS. (2024, 03 11). *Data Quality testing:*. Retrieved from Ways to test data validity and accuracy. Git For Data - lakeFS. : https://lakefs.io/data-quality/data-quality-testing/

Linkedin. (2023, 10 5). *How to Improve Data Quality with Effective Data Governance.* . Retrieved from Analytics8, Data & Analytics Consultancy. : https://www.linkedin.com/pulse/how-improve-data-quality-effective-governance-analytics8

linkedin. (2024, 02 18). *How can you define data quality roles and responsibilities?* Retrieved from linkedin: https://www.linkedin.com/advice/0/how-can-you-define-data-quality-roles-responsibilities-vswge

Ma, L. (2021, 09 17). *What is Data Management, actually?* Retrieved from DAMA-DMBOK Framework. Azure Data Ninjago & Dqops. : https://dataninjago.com/2021/09/15/what-is-data-management-actually-dama-dmbok-framework/

Martin, M. (2023, 12 30). *What is a Functional Requirement in Software Engineering?* . Retrieved from Guru99. : https://www.guru99.com/functional-requirement-specification-example.html

Microsoft Fabric. (2018, 04 9). *Custom Column using Power Query or DAX New column.* . Retrieved from Fabric community : https://community.fabric.microsoft.com/t5/Desktop/Custom-Column-using-Power-Query-or-DAX-New-column/m-p/392400

Mssaperla. (2024, 4 3). *Monitor metric tables - Azure Databricks.* . Retrieved from Microsoft Learn.: https://learn.microsoft.com/en-gb/azure/databricks/lakehouse-monitoring/monitor-output

Niels, R. (2021, 07 02). *ICT research methods* . Retrieved from ICT research methods : https://ictresearchmethods.nl/

Rabobank. (n.d.). *Data Management Consultant*. Retrieved from Rabobank: https://rabobank.jobs/en/job/data-management-consultant/JR_00082142/

Rachidi, L. H. (2024). *Data quality Management with DataBricks* . Retrieved from Databricks: https://www.databricks.com/discover/pages/data-quality-management

Rad, R. (2019, 6 3). *Create a Profiling Report in Power BI: Give the End User Information about the Data.* . Retrieved from RADACAD: https://radacad.com/create-a-profiling-report-in-power-bi-give-the-end-user-information-about-the-data

Rajeshsetlem. (2024, 01 11). *Failover groups overview & best practices - Azure SQL Database.* Retrieved from Microsoft Learn.: https://learn.microsoft.com/en-us/azure/azure-sql/database/failover-group-sql-db?view=azuresql

Richman, J. (2023, 06 21). *What is data quality?* Retrieved from Dimensions, standards, & examples. Estuary. : https://estuary.dev/data-quality/

Rome, P. (2024, 03 24). *Perforce*. Retrieved from What are Non Functional Requirements — With Examples. Perforce Software. : https://www.perforce.com/blog/alm/what-are-non-functional-requirements-examples

Segment. (2023, 03 08). *How to Prevent Data Discrepancies Before They Occur*. Retrieved from Twilio segment: https://segment.com/blog/data-discrepancy/

Sheldon, R. &. (2024, 03 04). *Data quality*. Retrieved from Data managment : https://www.techtarget.com/searchdatamanagement/definition/data-quality

Simplilearn. (2023, 11 17). *What is Requirement Analysis.* . Retrieved from Simplilearn: https://www.simplilearn.com/what-is-requirement-analysis-article

Singh, J. (2021, 10 22). *What is the Difference Between RPO and RTO? Druva Explains.* . Retrieved from Druva: https://www.druva.com/blog/understanding-rpo-and-rto

Sluzki, N. (2023, 8 30). *8 Data quality Monitoring Techniques & Metrics to watch.* . Retrieved from IBM Blog.: https://www.ibm.com/blog/8-data-quality-monitoring-techniques/

Twentynext. (2024, 01 15). *Overons*. Retrieved from Twentynext: https://twentynext.nl/over-ons/

Van Der Burg, M. (2023, 07 20). *Wat en voor wie is een elektronisch cliënten dossier (ECD)?* Retrieved from Nedap Healthcare: https://nedap-healthcare.com/ons-home/nieuws-ecd/

Versantvoort, J. (2024, 02 29). Data engineer interview. (R. alashabi, Interviewer)

Whiting, G. (2024, 03 23). *ExploreWMS*. Retrieved from How much WMS software costs and how to set your budget.: https://www.explorewms.com/how-much-wms-software-costs-and-how-to-set-your-budget.html

Appendices A: project plan

PROJECT PLAN

Data Health in Healthcare: A Quantum Leap for Quality Assurance and Automated Excellence.

Date completed: 09-04-2024

Author: Ramy Alashabi

Version: 1.7

Status: Final

9 APRIL 2024

Inhoud

Document History 55
Table of figures and tables 58
1. Introduction 60
1.1. Document objective 60
1.2. Reader's guide 60
2. Context & Background 61
2.1. About the client: 61
2.2. Project's current situation: 61
2.3. Problem description: 63
3. Project statement 63
3.1. Project goal 63
3.2. Project scope 63
3.3. Deliverables & non-deliverables 64
3.4. Work breakdown structure 64
3.5. Project constraints 65
3.6. Project risks 65
3.7. Methodology 66
3.8. Communication plan 67
3.8.1. Stakeholders/team 67
4. Research 68
4.1. Main research question 68
4.2. Research methods 68
5. Phase planning 71
5.1. MoScoW 71
5.2. Table of phases 72
5.3. Required skills 73
5.4. Project lead & product owner 75
6. Appendices 76
Appendices A: 76

Appendices B: 76
Appendices C: 77
Verwijzingen 77

Document History

Revisions

Version	Status	Date	Changes
0.1	Draft	20-02-2024	Create document
0.2	Draft	26-02-2024	Waiting for feedback
0.3	Draft	27-02-2024	Implemented verbal feedback.
1	Semi final	29-02-2024	Finished implementing verbal feedback
1.1	Final	11-03-2024	Finalized the PP with a concrete scope
1.2	Draft	14-03-2024	Processing canvas feedback
1.3	Draft	18-03-2024	Processing feedback form company visit.
1.4	Draft	22-03-2024	Processing extra feedback from data scientist with edits to research questions.
1.5	Final	25-03-2024	Implementing feedback on research questions from Rink
1.6	Final	04-04-2024	Implementing minor feedback from assessors and trimmed the chapters
1.7	Final.Final	09-04-2024	Reduced so much of chapter 2.2 and 2.3

Approval

This document requires following approvals:

Version	Date approval	Name	Function	Signed

Distribution

This document is distributed to:

Version	Date distribution	Name	Function
0.1	23-02-2024	Maikel	Company tutor
0.2	26-02-2024	Fontys	Tutor en 2nd assessor
1	29-02-2024	Fontys, Rink	Tutor en assessor

1.1	11-03-2024	Maikel en Rink	Begeleiders
1.2	14-03-2025	Maikel en Rink	Begeleiders
1.3	18-03-2024	Maikel en Rink en Ni	Begeleiders & data scientist
1.5	25-03-2024	Maikel en Rink	Begeleiders
1.6	04-04-2024	Maikel en Rink	Begeleiders
1.7	09-04-2024	Rink	begeleider

List of Terms & Abbreviations

Term	Meaning
Data integrity	Part of data quality as it focuses on the completeness, accuracy, uniqueness and consistency of the data.
deciphering	“to discover the meaning of something written badly or in a difficult or hidden way” (Cambridge, 2024)
leveraging	“power to influence people and get the results you want” (cambridge, 2024)
multifaceted	“Having many different parts or sides” (cambridge, 2024)
beacon	“a light or fire in a place that is easy to see, such as on the top of a hill, that acts as a warning or signal” (cambridge, 2024)
harness	“to control something, usually in order to use its power” (cambridge, 2024)
hindering	“to limit the ability of someone to do something, or to limit the development of something” (cambridge, 2024)
robust	“strong and unlikely to break or fail” (cambridge, 2024)
consolidating	“to add amounts or sets of numbers together to form a single amount, statement, etc” (cambridge, 2024)
bolstering	“to support or improve something or make it stronger” (cambridge, 2024)

Management summary

This project aims to enhance data integrity within healthcare organizations by implementing automated consistency checks, data quality metrics, and compliance checks in their data and reporting platform rollout. The primary objective is to ensure the accuracy, reliability, and uniqueness of healthcare data for its intended purposes. Poor data integrity in healthcare can lead to severe consequences, affecting patient care, regulatory compliance, research integrity, and stakeholder trust. By proactively addressing data quality issues, this project seeks to mitigate risks and improve overall healthcare operations.

Twentynext, the client for this project, specializes in deciphering big data and aims to assist organizations in extracting maximum value from vast datasets. Axion Continu, a client of Twentynext, seeks assistance in streamlining data sources from four different companies to enhance efficiency and accessibility. The proposed solution involves consolidating data sources into a single platform to address fragmentation and accessibility issues.

The project's goal is to elevate data integrity and reliability within Axion Continu's operations by implementing validation measures, defining quality standards, and establishing quality assurance protocol. Additionally, a dashboard will be developed to visualize errors for proactive management.

The project follows specific research methods that includes best practices analysis, literature study, document analysis, interviews, and guideline conformity analysis. These methods ensure a comprehensive understanding of data quality challenges in healthcare and inform evidence-based solutions tailored to industry standards and regulatory requirements.

The project follows an agile methodology, split into sprints of work to manage uncertainty, improve transparency, and shorten planning time. Regular meetings with stakeholders, including Twentynext and Fontys tutors, ensure project alignment and progress tracking. The communication plan involves meetings via Teams, emails, and personal interactions.

The project is guided by a clear timeline and deliverables, with a focus on addressing core questions and implementing effective data quality metrics. Risk mitigation strategies are in place to address potential challenges, including time constraints, scope creep, and communication issues.

In general, this project seeks to elevate the quality and reliability of healthcare data, ultimately improving patient care, regulatory compliance, and stakeholder trust. Through collaborative efforts and adherence to best practices, we aim to deliver a

solution that meets the diverse needs of healthcare organizations while ensuring accountability, transparency, and regulatory compliance.

Overall, Key aspects outlined in this project plan include the objective of the project, project structure, work schedule and deadlines, reason for the project, resources utilized, and methodology of work. The document serves as a foundational guide, facilitating effective project management, client approval, and progress monitoring. It also serves as an assessment tool for the student and provides a framework for future project endeavours.

Table of figures and tables

Figure 1 Introduction on the Medallion architecture, layers and their functions 8
Figure 2Showing The Learning Outcomes Of the Assessment Forum..... 10
Figure 3 Showing The Constraints Triangle..... 10
Figure 4 Showing the Sprint Way of Working 11
Figure 5 Work BreakDown Structure 19
Figure 6 Stakeholders description 21

Table 1 Describing What will and will not Be Delivered 9
Table 2 Describes The Potential Risks and Mitigations..... 11
Table 3 Contains Contact Information And Roles 12
Table 4 Elaboration On The Research Methods 13
Table 5 MoScoW On The Scope 15
Table 6 MoScoW on data analysis/ integrity dimensions 15

Table 7 MoScoW on the Data domains of AxionContinu
..... 15
Table 8 Shows the timing of the phases and the priority
..... 16
Table 9 Describes the Required Skills Based on The Learning Outcomes
..... 16

Introduction

Welcome to this project that is dedicated to improving data integrity in nursing homes organizations! My goal is to assess the present data integrity and provide guidance on addressing rising issues. I plan to achieve this by offering recommendations and conducting thorough automated consistency checks, as well as developing a comprehensive protocol for data quality metrics aimed at ETL engineers within the organization's data and reporting platform implementation. As according to the stakeholder, this project is vital because poor data quality can have serious consequences in nursing homes, impacting patient care, regulatory compliance, research integrity, and stakeholder trust, and this can be seen in Figure 6 .By addressing data integrity proactively, I aim to mitigate these risks and improve overall healthcare operations.

Document objective

This project plan for the "Data Quality" project aims to outline the following key aspects:

1. Objective of the Project
2. Project Structure
3. Work Schedule and Deadlines
4. Reason for the Project
5. Resources Utilized
6. Methodology of Work

This document serves as a foundational guide for the project, providing essential information to ensure project success, obtain client approval, monitor progress, and facilitate effective project management. For the student, it serves as a basis for assessment and demonstration of project management skills, while for the company, it provides a solid framework for current and future projects.

Reader’s guide

In this document, there are 6 chapters and below is a brief overview of what to expect in each section:

1. Introduction: Provides an overview of the project's objectives and the importance of addressing data quality issues in healthcare organizations.
2. Context & Background: Offers insights into the client, Twentynext, and the specific challenges faced by Axion Continu in their data management process. It outlines the current workflow and identifies key areas for improvement.
3. Project Statement: Defines the project goal, scope, and deliverables. It also outlines the constraints, risks, and methodology adopted for project execution.

4. Research: Details the main research questions and methodologies used to address them. It provides a breakdown of research methods applied to each research question.
5. Phase Planning: Outlines the timeline and deliverables for each phase of the project, including research phases and release phases. It also highlights the soft and hard skills necessary for project success, including critical thinking, communication, coding proficiency, and project management. Additionally, it describes the roles and responsibilities of the project lead (student) and the product owner (Twentynext) throughout the project lifecycle.
6. References: it's a list to all the referenced names and methods.

This document serves as a roadmap for understanding the project's objectives, methodologies,
and deliverables. It provides stakeholders with valuable insights into the project's progress and ensures alignment with client expectations.

Context & Background

About the client:

The purpose of this chapter is to talks about the client and what they do.

In a complex world in which data is increasing exponentially, Twentynext helps organizations create value from data and make relevant information available at the right time. With data science and artificial intelligence experiences coming from Twentynext, it enables organizations to make better data-driven decisions, gain more control and improve results.

If organizations do not have the people to set up a complex (data) project themselves, Twentynext can start a project with the organization and deploy the right people (Twentynext, 2024). Twentynext can also support in managing the organizations data and AI solutions. Therefore, imagine an organization that does not (temporarily) have the right expertise in their own data science team, this opens the path for Twentynext to collaborate and support the organization with their skilled and experienced interim professionals.

One of Twentynext clients is AxionContinu. AxionContinu is a nursing home organization that is founded in 2013 and is based in Utrecht. Additionally, they provides essential care for elderly citizens and patients in need (AxionContinu, 2023). They approached Twentynext with an IT issue that will be further detailed in the following chapter.

Project's current situation:

The purpose of this chapter is to describe the current situation of the project within Twentynext in regards to AxionContinu in general.

AxionContinu, a client of Twentynext, seeks assistance to address an issue in their current process. Currently, data and dashboards are sourced from four different companies, leading to confusion and inconsistencies. These 4 companies are :

1. Infent
2. Cognos
3. OGD
4. MijnCaress

Infent typically operates within the VISMA dashboard, showcasing financial data linked to mijncaress.

Cognos, OGD, and Mijncaress collaborate on mijncaress data encompassing patient information, schedules, and more. This old fragmented/ dependant approach by
AxionContinu resulted in missing data and delays, hindering Axion Continu's ability to access complete information necessary for their operations. However, the proposed solution by Twentynext involves streamlining data sources from four to one. Progress has been made by collecting data from Software as a Service (SaaS) platforms and depositing it into an SQL lake as an initial step towards consolidating data sources.

After the data collection phase, Twentynext utilizes Databricks for ETL operations, adhering to the Medallion Lakehouse architecture shown in Figure 1. This architecture, represented by bronze, silver, and gold layers, serves to systematically enhance data quality as it progresses through each layer. Each layer within this architecture fulfils a unique function, as illustrated below.

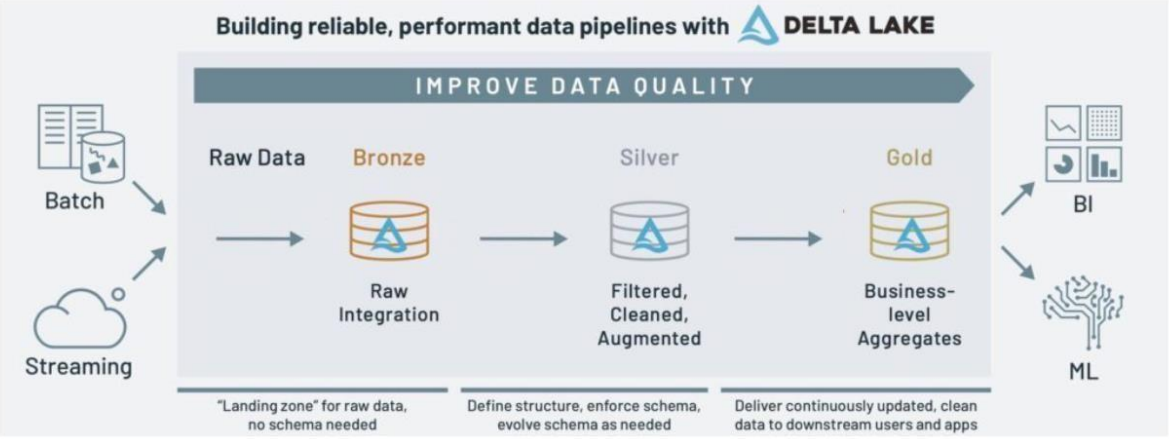


Figure 1 Introduction on the Medallion architecture, layers and their functions

Within the medallion architecture, bronze involves raw data collection and integration with external sources, while silver focuses on structuring,

cleansing, and augmenting the data. Gold entails constructing data models with factual and dimensional tables, representing business-level aggregates. Currently, the project is transitioning between the bronze and silver layers, and no actions have been taken regarding data quality.

According to a colleague, the importance of this project is crucial, especially given the transition between layers (Versantvoort, 2024). It represents the essential next step in ensuring data quality, upon which all subsequent actions rely.

Problem description:

The purpose of this chapter is to outline the current challenges faced by Axion Continu, a client of Twentynext, particularly within the context of data management.

At present, the focus lies on ensuring data quality and integrity during the rollout of the healthcare organization's data and reporting platform, specifically within the transition from the bronze to silver layers. Risks during this phase, such as missing data and field/table dropping, threaten the reliability of the data. Additionally, AxionContinu relies on data and dashboards from multiple sources, leading to various operational and reputational risks, according to one of the stakeholders in Figure 6. Although consolidating data sources into a single platform presents an opportunity for improvement, this aspect falls outside the scope of the current project. Nevertheless, addressing challenges like data format discrepancies and data loss risks between the bronze and silver layers is crucial for improving data integrity, which is the main focus of this project. A simple illustrative instance showcasing data integrity is envisioning a scenario where a row in the gender column is either missing for a patient or contains a "?" symbol, or encountering a BSN number exceeding 9 characters.

Project statement

Project goal

The goal of this project is to implement validation measures on existing data, establish accepted data integrity standards, and develop a dashboard for visualizing errors to enable and increase the efficiency of data engineers in the ETL process.

Project scope

The project will focus on improving data integrity at AxionContinu by implementing validation measures, defining quality standards and quality

assurance protocol. A dashboard/notebook will be developed to visualize errors including general data analysis and consistency. Additionally, the scope is limited to analysing the data integrity (consistency) of one specific sources (Mijncaress) within the "bronze-silver A" layers. Anything beyond this falls out of this scope. In the case of extra time, the scope will increase. Refer to Chapter 5 for detailed scope phases and priorities.

Deliverables & non-deliverables

Table 1 Describing What will and will not Be Delivered

The project includes:	The project does not include:
1. Research document/protocol/advice	1. Data collection between platforms
2. Analysis on current situation	2.Contact client outside the organization
3. Dashboard/ notebook	3.Improving Master Data Management & data quality of software solutions, e.g. AFAS, Visma, MijnCaress.
4. Thesis report	

Work breakdown structure

The work breakdown structure can be found in chapter 6 appendices.

The project's success relies heavily on addressing core questions, identifying current situation’s bottlenecks , and implementing effective data quality metrics. To my assumption, It's essential for key stakeholders to thoroughly evaluate the final product before delivery, as this is an essential step for the subsequent actions. Furthermore, both the client and the school's tutor will assess the project during the concluding demonstration or final company visit, based on the assessment dimensions that are provided by fontys which can be seen in the figure below.

U: Unsatisfactory/Onvoldoende, S: Satisfactory/Voldoende, G: Good/Goed, O: Outstanding/Uitmuntend

	Assessment dimensions	U	S	G	O	Feedback
1	Professional Duties					
2	Situation-Orientation					
3	Future-Oriented Organisation					
4	Investigative Problem Solving					
5	Personal Leadership					
6	Targeted Interaction					

Figure 2 Showing The Learning Outcomes Of the Assessment Forum

Project constraints

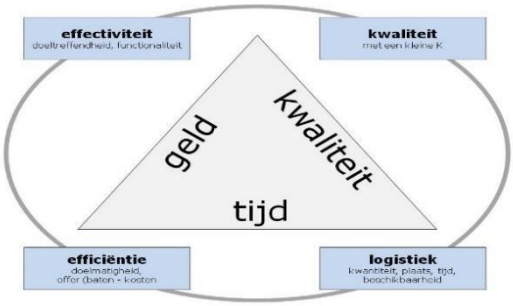


Figure 3 Showing The Constraints Triangle

This project too might face some constraints as mentioned below:

- Time constraint, as only within 18 weeks does this project must be in the final working phase
- Quality and time, as the student will be working on the project independently therefore time>quality .
- As this project is for institutional training for the student money constraint is not much of a constraint to the project

Project risks

Table 2 Describes The Potential Risks and Mitigations

Risk	*Probablity*	*Impact*	*prevention*

Sick leaves	*50%*	*100%*	-
Vacations	*50%*	*50%*	*Make a good plan*
Insufficient time	*50%*	*50%*	*Make a realistic plan*
Change of approach	*50%*	*100%*	*Stick to the plan and if any changes make it small*
Insufficient data	*50%*	*40%*	*More research and data collection*
Insufficient communications	*50%*	*100%*	*1xper week meeting with couch at school and twice per week with tutor at work*
End product does not meet expectations	*50%*	*100%*	*Keep the client in the loop always with constant feedbacks*
Difference between ICT and Accountant mindset	*50%*	*100%*	*Keep a clear communication and run the numbers by them first*
Scope Creep/ expands	*50%*	*100%*	establish clear project objectives and deliverables upfront and maintain regular communication with stakeholders to manage expectations.
Project detriment	*50%*	*100%*	Ensure the availability of the data and clear communications.

Methodology

The approach that will be used is agile (Agile alliance, 2022), and that is because agile is split into sprints or small steps of work. This causes the team that works with agile to manage uncertainty more, improves on transparency by keeping a back log and it also shortens the planning time. This methodology will be used as following:

1. Sprint meeting with the company tutor twice a week
2. Sprint meeting with the school tutor once every Bi-week
3. Backlogs tracking and roadmap
4. Clients satisfaction through continuous delivery
5. 2 weeks are given per sprint.
6. Demo will be shown every 2nd sprint or every sprint for constant feedback 7. Retrospective is to be set at the end of the 17 weeks

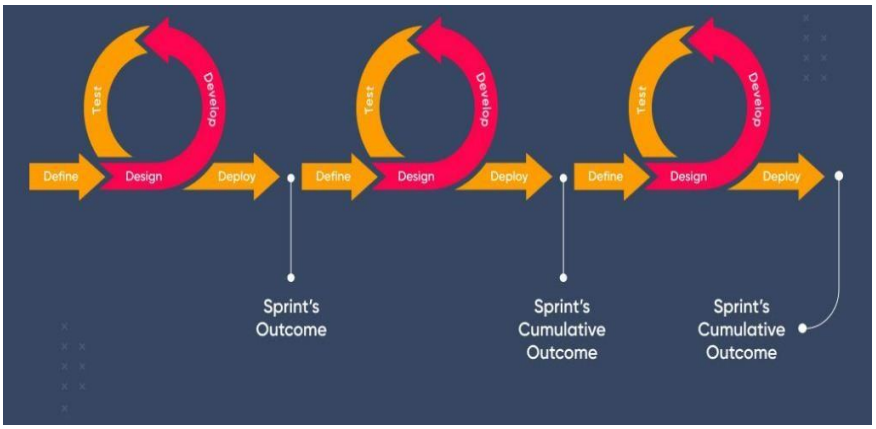


Figure 4 Showing the Sprint Way of Working

Communication plan

All communications are done via teams, personally, or Emails.

1. Meeting the company tutor is done in the office twice per week.
2. Meeting with the school tutor is done every other week online via teams or on location.
3. Using emails to schedule a personal meeting with the other stakeholders when needed.

Stakeholders/team

Within the company of Twentynext the team in total consist of 20 personnel. In this project the team's/ stakeholders information can be seen in the table below:

Table 3 Contains Contact Information And Roles

Name +e-mail	Abbr.	Role/tasks	Availability
Ramy Fuad Email:r.alashabi@student.fontys.nl	DE/Student	*Business and data analyst /Carries out the whole internship project*	*During the whole phases.*
Maikel Maasakkers Email: m.maasakkers@twentynext.nl	DA/owner	*Business and data analyst/ Tutor at the company*	*Twice every week in every phase*
Rink Lycklama Email:rink.lycklama@fontys.nl	Docent/tutor	Guide and tutor throughout the internship	Once every 2nd week in every phase

Research

Main research question

In here the research questions will be listed below, to breakdown the main problems into 1 main question that is followed by sub questions.

How can a methodical instrument be designed that effectively measures data integrity, incorporating flexible parameters aligned with business rules, while ensuring consistency and reliability in the assessment process?

a. How can the development of a protocol ensures the integrity of the data (complete, accurate, consistent and unique) within nursing home data systems?
Expected outcome/product: a protocol document that advises the data engineer on what to do with examples and actions to ensure data quality during ETL phase.

b. What metrics or code of validation measures should be taken into account for ensuring data integrity, and can they be tailored for diverse business requirements ?
Expected outcome/product: Code lines or metrics that measures the data integrity and an analysis document/markdown for the current data situation with noticed findings and patterns with testcases.

c. How can a monitoring dashboard be developed, incorporating seamless integration in the IT landscape, while measuring the dimensions of data integrity and employing suitable visualizations to enhance clarity and insight?
Expected outcome/product: A prototype of a Power BI dashboard /notebook that visualize the dimensions of data integrity.

Research methods

Research methods are activities, tactics and procedures used in the compendium of data or proof for analysis with the purpose of revealing additional information or establish a vivid understanding of the topic. As it can be seen in the table below, there are plenty different types of research methods and for this project we are using the methods provided by Fontys "ICT research methods."

Table 4 Elaboration On The Research Methods

Question	Method	How?	Why?

How can the development of a protocol ensures the integrity of the data (complete, accurate, consistent and unique) within nursing home data systems?		**Best good and bad practices**: to gain insights from existing data quality frameworks in healthcare, ensuring a comprehensive approach to designing a metric that aligns with industry standards and regulatory requirements. This method facilitates the identification and incorporation of proven practices, optimizing the effectiveness and credibility of the data quality metric for healthcare data. **Literature study**: guidance and best practices from existing literature to inform the development of the	**Best good and bad practices**: Because this method allows for a comprehensive review of existing data quality frameworks in healthcare. **Literature study**: Because it draws from established research, ensuring evidence-based design. **Document analysis**: By learning from past implementations, it provides

		data quality metric and develop a search plan, identify relevant keywords, and evaluate material to select pertinent resources. **Document analysis**: analyse the documents available to study what was previously done for the data quality by the previous company. **Interview** : To collect expertise on how to deal with data quality and get clear definition on data quality in the health department. **Guideline conformity analysis**: ensure adherence to established standards and guidelines in designing the data quality metric for healthcare.	insights for improvement and innovation. **Interviews**: Because gathering firsthand expertise ensures alignment with healthcare context and specific needs. **Guideline conformity analysis**: Ensures compliance with industry standards, mitigating risks and ensuring quality assurance.

What metrics or code of validation measures should be taken into account for ensuring data integrity, and can they be tailored for business requirements?		**Literature study**: to identify existing research, frameworks, and methodologies related to automated consistency checks in healthcare data management. This allows for understanding key components such as data validation rules, anomaly detection techniques, and error handling strategies. **Data quality check** : to ensure the adequacy of data quality for subsequent analyses in healthcare settings. Develop robust test cases and automate them into test scripts to detect incomplete, incorrect, or inaccurate data. **Ethical check** : to ensure that automated consistency checks in healthcare data management align with ethical norms and values. Investigate potential ethical dilemmas that may arise from the implementation of automated checks and consider diverse perspectives to inform decision-making. **Interview**: Conduct interviews with representative participants, probing their opinions, behaviours, and experiences related to data quality management.	**Literature study**: Because it draws from established research, ensuring evidence-based design. **Data quality check:** because it offers a practical validation of data quality, ensuring reliability and accuracy for subsequent analyses. **Ethical check:** Because it ensures alignment with ethical norms, mitigating potential ethical dilemmas and promoting responsible data management. **Interviews**: Because gathering firsthand expertise ensures alignment with healthcare context and specific needs.
How can a monitoring dashboard be developed, incorporating seamless integration in the IT landscape, while measuring the dimensions of data integrity and employing suitable visualizations to enhance clarity and insight?		**Community research** :Research stack overflow or BI community for the best ways to visualize the data quality. **Prototyping**: Create a small PoC that would be able to visualize the errors and data quality based on the components of the automated checks. **Peer review** : Organize a peer review session with a representative group of colleagues and external experts to evaluate the proposed governance strategies in a structured manner.	**Community research:** Because utilizing platforms like Stack Overflow or BI communities enables access to a wide range of expertise and perspectives. **Prototyping:** Because creating a Proof of Concept (PoC) allows for practical testing and validation of the proposed dashboard's functionalities. **Peer review:** Because it provides structured feedback for accountability and regulatory compliance from a co-worker.

Phase planning

MoScoW

Given the substantial scope of the project, it is advisable to employ a MoSCoW table to effectively slice the project's components within the internship's 17-week timeline. However, it's worth noting that the application of MoSCoW may not directly translate to real-world scenarios, where all aspects are typically deemed significant in the business.

Table 5 MoScoW On The Scope

Scope	Must	Should	Could	Wont
1. Implementing validation methods: Validating data consistency, completeness and etc	X			
2. Defining quality standards: Includes guidelines, benchmarks, criteria and patterns of data	X			
3.Establishing quality assurance mechanism: Creating a process or system to maintain the quality like audits, checks and reviews			X	
4.Develop a prototype dashboard/ notebook: to visualize the errors and the metrics of the passing through data as a proof of concept	X			
5.Develop a dashboard for the production phase: visualize errors and metrics but for all data				X

However, If the essential tasks designated as "Musts" are completed ahead of schedule, the student may proceed to address the objectives categorized as "Shoulds" within the scope, followed by those categorized as "Coulds."

Table 6 MoScoW on data analysis/ integrity dimensions

Data Integrity dimensions	Must	Should	Could	Wont
Consistency	X			
Accuracy		X		
Uniqueness				X
Completeness			X	
General analysis	X			

Table 7 MoScoW on the Data domains of AxionContinu

Data domains	Must	Should	Could	Wont

Financieel		X		
Kwaliteit				X
Personeel		X		
Productie		X		
Rooster			X	
Patiënt	X			

Explanation on the data domains and their locations:

1. **Financieel**: Bedrijf, kosten, groot boek en etc. (locaton: visma.net)
2. **Patiënt data**: age, name en etc. (location: mijncaress) 3. **Kwaliteit**: Client information. (Location: All applications) 4. Personeel: medewerker, salaris, functie en etc (Location: afas)
5. **Productie**: afdeling, factuur, product en etc (Location: mijncaress)
6. **Rooster**: Activitiet, functie, dienstverband en etc (Location: mijncaress)

Table of phases

The table below shows the phases planning of the documents research and release phases.

Table 8 Shows the timing of the phases and the priority

Phase	Start/end date	Deliverables (MoScoW)
Research phase	19-02-2024/04-03-2024 (2 weeks)	Draft Version of project plan. (M) Research document beginning. (M)
Research phase	05-03-2024/ 19-03-2024(2 weeks)	Research document Draft.(M) Draft Version of project plan. (M)
Research phase	20-03-2024/ 05-04-2024(2weeks)	research document Draft.(M) Draft Version of project plan. (M)
Research phase	06-04-2024/20-04-2024 (2weeks)	Final research document. (M) Draft Version of project plan. (M)
1st release phase	21-04-2024/ 05-05-2024(2 weeks)	Draft Version of project plan. (M) Thesis beginning.(M)

2nd release phase	06-05-2024/20-05-2024(2 weeks)	Thesis draft.(M) Dashboard beginning.(C)
3rd release phase	21-05-2024/ 04-06-2024(2 weeks)	Thesis draft.(M) Dashboard draft.(C)
4th release/ semi final phase	05-06-2024/18-06-2024(2 weeks) Final date.	Final thesis and submission.(M) Final dashboard.(C)
Final release phase	20-06-2024/ 27-06-2024(1 week)	Final edits of the dashboard and document.(S) Preparation of presentation.(M)

Required skills

There are skills required for this project and the learning outcomes and they are as following:

Table 9 Describes the Required Skills Based on The Learning Outcomes

Assessment dimensions	Soft Skills required	Hard skills required
1. professional duties	adaptability: being able to adapt easily to changes. Attention to details: precision in executing duties.	SDLC knowledge: knowing the phases of analysis, design, realize and etc. Domain-specific proficiency: Mastery in an IT domain, knowledge of relevant tools and technology.
	3. Time management: allocating time and resources to meet deadlines.	3. Architectural layers expert: competence in navigating and integrating different layers per project requirements.
2. Situation orientation	Problem sensitivity: being able to recognize and predict potential issues. Adaptation: adapting existing knowledge and skills to a new situation. Value creation: focus on creating a valuable result and in this case data quality metrics.	Context analysis: being able to analyse the current situation and context technology analysis: being able to recognize and use new technology. ethical analysis: being able to recognize what is ethical and what is not in personal data.

3.Futureoriented organisation	Strategic thinking: Analyse long term trends and future developments. Future perspective: envision future scenarios and asses their potential impact on the outcome. Stakeholder engagement: Engaging stakeholders to identify business legitimization, values and ethics, long term.	Business analysis: being able to analyzie the business requirements and opportunities. Sustainability integration: Incorporate sustainability considerations in the project. Ethical decision making: Being able to recognize ethical dilemmas and uphold ethical principles.
4.Invetigative problem solving	1. Critical thinking: Analyse complex issues, identify cause and develop solution. 2. Problem solving mindset: Positive attitude in tackling challenges, explore alternative approaches and seek continuous improvement. 3. Research methodology: design and conduct research, problem analysis, hypothesis formulation and etc.	Root analysis: being able to identify the root of the challenge. Research design: designing research methodologies, selecting appropriate data collection techniques and ensuring validity and reliability. validation techniques: assessing the effectiveness and suitability of the proposed solution.
5.Personal leadership	Self-reflection: Being able to reflect on my strengths and weaknesses in area of growth. Initiative: Proactive approach in taking ownership of the project. 3. Visionary thinking: Ability to envision future career aspirations and align personal development goals with professional objectives.	Project management: proficient in project planning, execution and etc. Feedback solicitation and integration: Skilled in seeking constructive feedback from surrounding and integrate them. Career planning: knowledge in career pathway and opportunities in the IT industry.
6.Targeted interaction	1.communication: being able to communicate vividly and diplomatically. 2. Cultural sensitivity: awareness of cultural differences and norms	1. Communication planning: develop a communication plan to ensure effective and timely information sharing. 2. Presentation skills: being able to present an engaging and informative presentation of the project.

	and adapting to it. 3. Conflict resolution: being able to understand and to explain different point of views when conflict rises.	

Project lead & product owner

Throughout the development process, various milestones will be reached, each marked by meetings with stakeholders including the client, represented by Twentynext, and the school tutor from Fontys. Feedback gathered during these meetings will guide project progression. Twentynext and the Fontys tutor will jointly assess the project before its final publication. The student leading the project will oversee its execution.

Appendices

Appendices A:

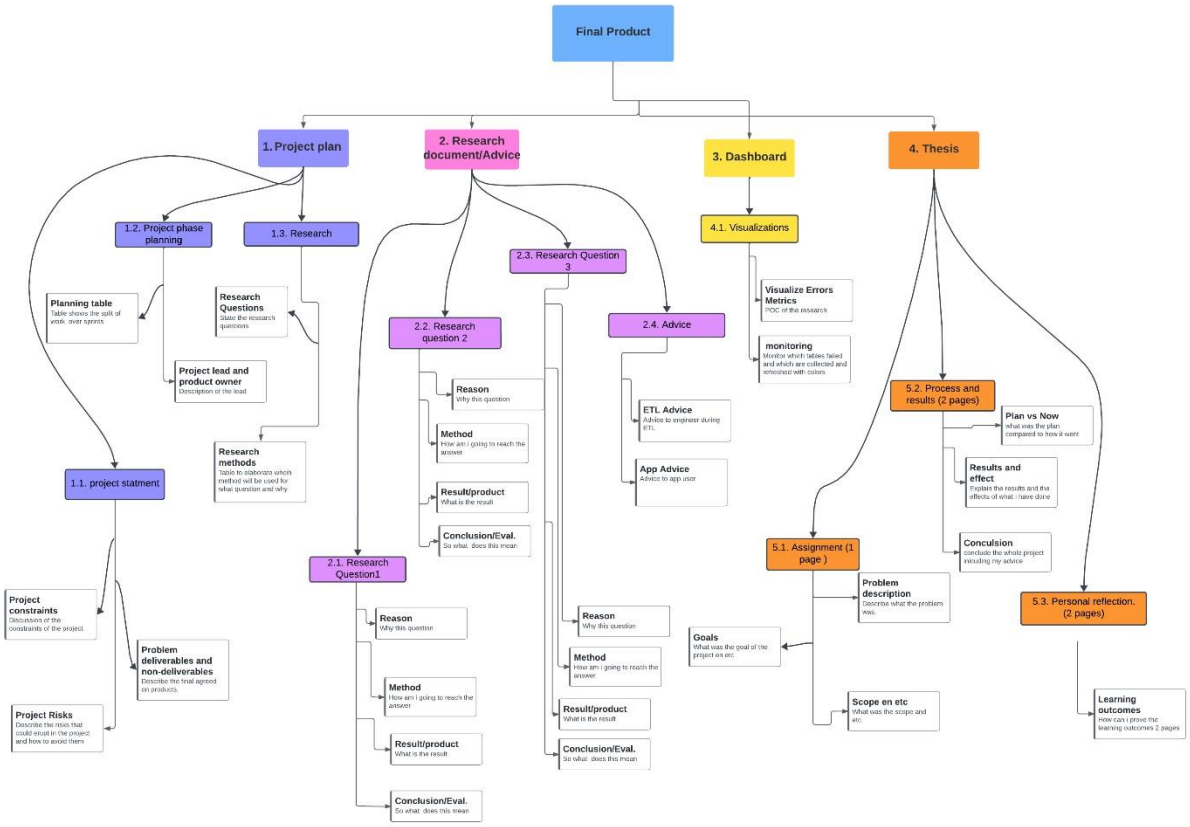
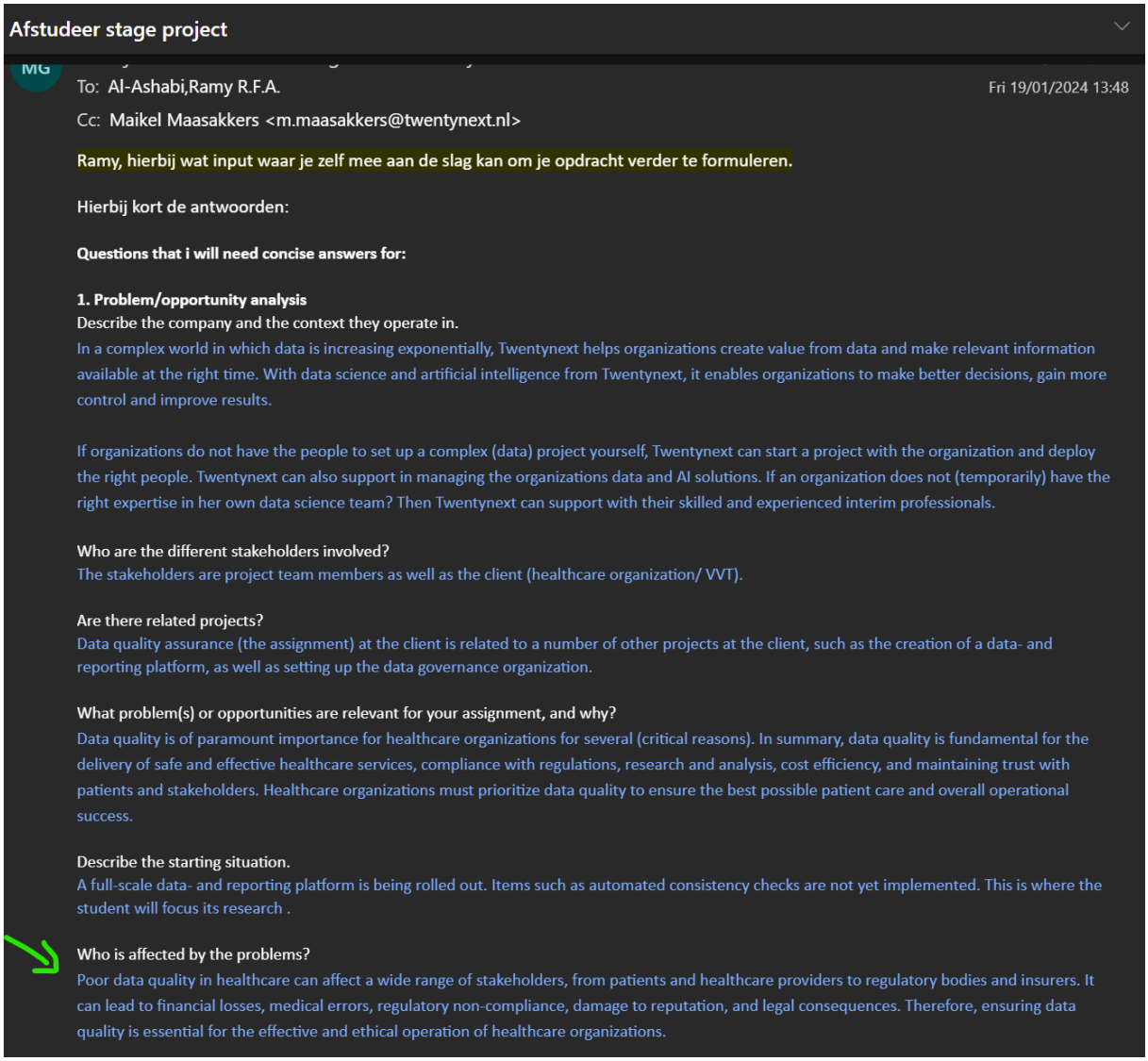


Figure 5 Work BreakDown Structure

Appendices B:

Transcript Jasper.docx

Appendices C:



Afstudeer stage project

MG

To: Al-Ashabi,Ramy R.F.A. Fri 19/01/2024 13:48

Cc: Maikel Maasakkers <m.maasakkers@twentynext.nl>

Ramy, hierbij wat input waar je zelf mee aan de slag kan om je opdracht verder te formuleren.

Hierbij kort de antwoorden:

Questions that i will need concise answers for:

1. Problem/opportunity analysis

Describe the company and the context they operate in.

In a complex world in which data is increasing exponentially, Twentynext helps organizations create value from data and make relevant information available at the right time. With data science and artificial intelligence from Twentynext, it enables organizations to make better decisions, gain more control and improve results.

If organizations do not have the people to set up a complex (data) project yourself, Twentynext can start a project with the organization and deploy the right people. Twentynext can also support in managing the organizations data and AI solutions. If an organization does not (temporarily) have the right expertise in her own data science team? Then Twentynext can support with their skilled and experienced interim professionals.

Who are the different stakeholders involved?

The stakeholders are project team members as well as the client (healthcare organization/ VVT).

Are there related projects?

Data quality assurance (the assignment) at the client is related to a number of other projects at the client, such as the creation of a data- and reporting platform, as well as setting up the data governance organization.

What problem(s) or opportunities are relevant for your assignment, and why?

Data quality is of paramount importance for healthcare organizations for several (critical reasons). In summary, data quality is fundamental for the delivery of safe and effective healthcare services, compliance with regulations, research and analysis, cost efficiency, and maintaining trust with patients and stakeholders. Healthcare organizations must prioritize data quality to ensure the best possible patient care and overall operational success.

Describe the starting situation.

A full-scale data- and reporting platform is being rolled out. Items such as automated consistency checks are not yet implemented. This is where the student will focus its research .

Who is affected by the problems?

Poor data quality in healthcare can affect a wide range of stakeholders, from patients and healthcare providers to regulatory bodies and insurers. It can lead to financial losses, medical errors, regulatory non-compliance, damage to reputation, and legal consequences. Therefore, ensuring data quality is essential for the effective and ethical operation of healthcare organizations.

Figure 6 Stakeholders description

Verwijzingen

Agile alliance. (2022, 05 26). *What is Agile Software Development?* Retrieved from Agile alllinace:

https://www.agilealliance.org/agile101/

AxionContinu. (2023, 12 4). *Over ons.* . Retrieved from AxionContinu:

https://www.AxionContinu.nl/over-ons#actueel

cambridge. (2024). *beacon*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/beacon

cambridge. (2024). *bolstering*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/bolster?q=bolstering

cambridge. (2024). *consolidating*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/consolidate?q=consolida ting

Cambridge. (2024). *decipher*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/decipher

cambridge. (2024). *harness*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/harness

cambridge. (2024). *hinder*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/hinder?q=hindering+

cambridge. (2024). *leverage*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/leverage

cambridge. (2024). *multifaceted*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/multifaceted

cambridge. (2024). *robust*. Retrieved from cambridge: https://dictionary.cambridge.org/dictionary/english/robust

Niels, R. (2021, 07 02). *ICT research methods* . Retrieved from ICT research methods : https://ictresearchmethods.nl/

Twentynext. (2024, 01 15). *Overons*. Retrieved from Twentynext: https://twentynext.nl/over-ons/

Versantvoort, J. (2024, 02 29). Data engineer interview. (R. alashabi, Interviewer)

Appendices B: research document

Research/Advice document

Data Vigilance: Advancing Integrity Assurance and Automated Excellence

REMY ASHABI

Inhoud

1. Document goal 82
2. Requirement analysis 82
    2.1. What is requirement analysis? 82
    2.2. Business requirements 83
    2.3. User requirements 84
    2.4. System requirements 84
        2.4.1. Functional requirements 85
        2.4.2. Non-functional requirements 85
3. Research Question 1 86
    3.1. Reason 86
    3.2. Method 86
    3.3. Result/product 87
    3.4. Conclusion/ Evaluation 102
4. Research Question 2 104
    4.1. Reason 104
    4.2. Method 104
    4.3. Result/product 104
    4.4. Conclusion/Evaluation 107
5. Research Question 3 108
    5.1. Reason 108
    5.2. Method 109
    5.3. Result/product 109
    5.4. Conclusion 116

6. Advice/recommendation 117
6.1. Previous situation 118
6.2. Advised situation 118
7. References 119
8. Appendices 123
8.1. Appendices A: Align dumpstore 123
8.2. Appendices B: Kubus 124
8.3. Appendices C: Align infostore 124
8.4. Appendices D: Interview NI Kang 125
8.5. Appendices E: Interview Linda 139
8.6. Appendices F: Interview Rink 177
8.7. Appendices G: Advised protocol 214
8.8. Appendices H: Jupyter notebook screenshot 232
8.9. Appendices I: Interview Linda2 232
8.10. Appendices J: Interview Jasper 270

Figure 1 Requirements Pyramid 5
Figure 2 Dot framework methods used in this question .....
8
Figure 3 Data lifecycle 10
Figure 4 DMBOK Data governance diagram 12
Figure 5 DMBOK Pyramid 13
Figure 6 Dumpstore doc 15
Figure 7 Screenshot on infostore and the part that was helpful .....
16
Figure 8 Dot framework methods used for this question .....
27
Figure 9 Part of the protocol made 28
Figure 10 Excel that shows the rules and the checks per column .....
28
Figure 11 Example of errors found during analysis .....
29 Figure 12 Excel sheet that contains the patterns found the rules and what to check and
how ..... 30 Figure 13
Dot framework methods used for this question. 32
Figure 14 showing how to connect to databricks. 32
Figure 15 showing the manual code that was made manually. .....
33
Figure 16 Showing the code found after the research. .....
33
Figure 17 Semantic model 34
Figure 18 EDA made in Power BI 34
Figure 19 Home navigation page for the dashboard 34
Figure 20 BSN invalid struggle 35
Figure 21 Stakeholder requirments in a diagram 36
Figure 22 Improved visuals based on requirements 37
Figure 23 Improved drill-through visual based on requirements .....
37
Figure 24 showing the amount of patients per error. .....
38
Figure 25 Filter button closed 38
Figure 26 Filter page open 39
Figure 27 Applied filters without taking space 39
Figure 28 Document 1 provided by AxionContinu 46
Figure 29 Document 2 provided by axionconitnu 47
Figure 30 Document 3 provided by AxionContinu 47

Figure 31 Screenshot of the notebook made for the metrics 160

1. Document goal

The goal of this document is to document all the requirements (system or user requirements) and the research findings for all the research questions. Where the reason behind the question will be elaborated, the methods used while and to reach your research, the results of the research and what does it mean in regard to the project, will also be elaborated. Each chapter in this document is a question and each question has sub chapter.

NOTE: This document follows the same structure as the thesis document. Every research question has its own chapter. The 5th chapter might seem different than the thesis document, due to further research was recorded in the thesis. Additionally, Chapter 2 is a method that is used in chapter 5. However, it was added here instead of it having its own document.

2. Requirement analysis

Some information in this chapter is based on the intern's own research from 2 years ago, however updated.

2.1. What is requirement analysis?

Requirement analysis is a process where the collection of needs and expectation of a product is made. It is made based on an intensive communication between the stake holder and the software engineer/creator (Simplilearn, 2023). The purpose of requirement analysis is to accurately represent the stake holder's wishes, that can be translated into code and when made it meets the stakeholder's expectation. There are different types of requirements, however for this project we are analyzing the business requirements, user requirements, functional and nonfunctional requirements.

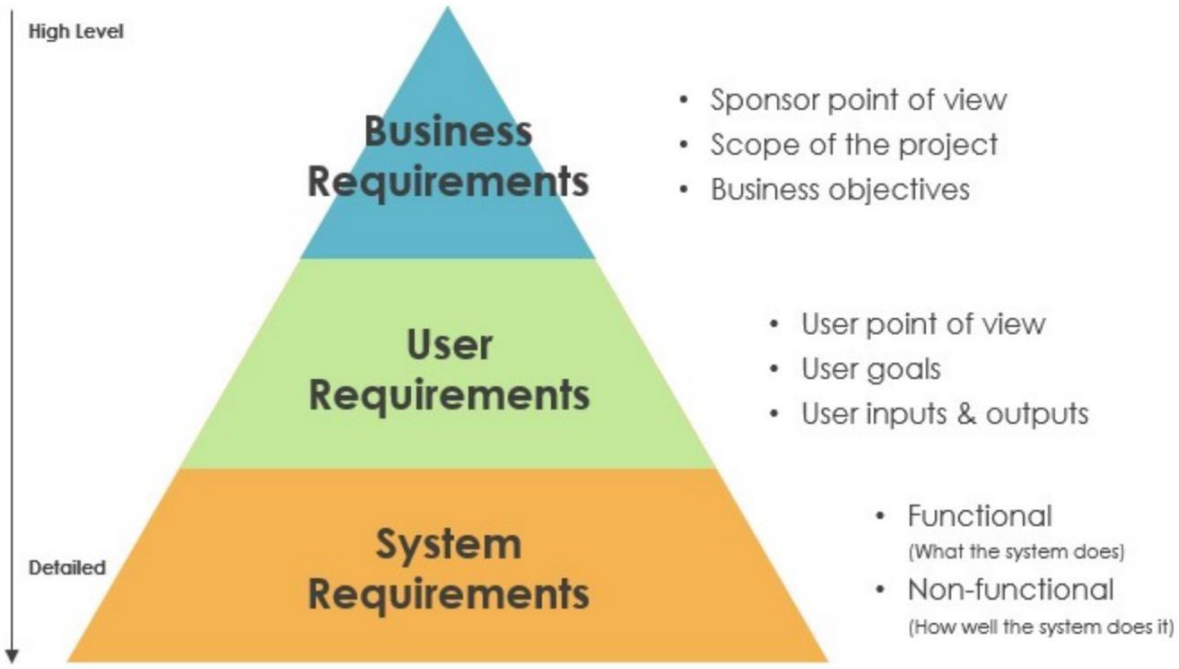


Figure 1 Requirements Pyramid

These requirements mentioned above, does not have to have 1 source, as it can be collected from a direct requirements interview, or a case/ process of work which will need to be translated to requirement, or even from researching on the internet (Aziz, 2023). In this project, the requirements are collected from all conducted interviews and research done, as in there was no specific interview that was made only for requirements collections, but mostly translated from the weekly 1-on 1 meetings with the mentor into requirements and user requirements.

2.2. Business requirements

Business requirements is basically not what a system must do, but it is what a business must do (i.e., Complete a process), in order to stay relevant (Cox A. , 2017). The purpose of this, is to make the business more efficient by understanding what needs to be done, and that is done by aliening the goals and visions of the company and intensive communication.
Business requirements:

1. BR01: AxionContinu aims to fortify its data governance framework by implementing procedures to maintain data integrity and quality throughout the ETL process.

2. BR02: AxionContinu seeks to enhance its data integrity assurance practices by introducing proactive measures for integrity validation and continuous monitoring of the current data repository.

2.3. User requirements

User requirements are basically requirements set by the user. It is basically a set of requirements that tells how the user wants to react with the system and what do they expect the system to do (Doel, 2018). User requirements act as a concrete base for the system as they have to be unique, concise and simple, and based on 1 user requirement multiple functional or non-functional requirements could be made.

User requirements:

UR01: User wants to see the number of errors per column in percentage.

UR02: User wants to see the number of errors per patient.

UR03: User wants to see the number of nulls per column.

UR04: User wants to see the consistency of Ucase per column.

UR05: User wants to see the value of the error.

UR06: User wants to see the remarks about the error.

UR07: User wants to see the input time of each error.

UR08: User wants to see the number of patients per error.

2.4. System requirements

System requirements are basically a set of rules that the system must contain, in order for the system to run smoothly as requested. However, it describes the behavior of the system and its functions, which makes it easier for the data analyst to build and stakeholder to understand (Cox A. , 2024) .There are 2 types of system requirements, Functional and non-functional requirements. The functional requirement is basically a description of what the system should do, its components or the behavior of the system. Mainly, what a system should show or display (Martin, 2023). The non-functional requirement is basically a constraint of how the system should display the functional requirements (i.e., speed of a graph to show, or speed of data load). It has mainly to do with the attributes of the functional requirements (Rome, 2024).

2.4.1. Functional requirements

SR01: System must be able to automatically show updated data.

SR02: System must be able to automatically validate (BSN -11proef) and display the new data from the database into dashboard.

SR03: System must be able to calculate and display the percentage of errors for each column within the dataset.

SR04: System must be able to provide a number of errors associated with each patient.

SR05: System must be able to count and display the number of null values for each column in the dataset.

SR06: System must be able to analyze and display the consistency of uppercase usage within each column.

SR07: System must be able to display the specific value of each detected error.

SR08: System must be able to provide detailed remarks for each detected error.

SR09: System must be able to record and display the input time of each detected error.

SR10: System must display the number of patients affected by each type of error in the dataset.

2.4.2. Non-functional requirements

NFR01: System must be able to display updated data within 6 seconds of data retrieval from the database to ensure real-time monitoring (related to SR01 and SR02).

NFR02: System must be capable of handling datasets with up to 1 million records without performance degradation to support future growth (related to SR03, SR04, SR05, SR06, SR07, SR08, SR09, and SR10)

NFR03: System interface must be intuitive and user-friendly, requiring no more than 2 hours of training for a new user to become proficient (related to all SRs).

NFR04: System must ensure that data validation processes (e.g., BSN - 11proef) have an error rate of less than 0.01% as an example to maintain data integrity (related to SR02). Depends on the requirements of AxionContinu.

NFR05: System must be compatible with major web browsers (Chrome, Firefox,
Safari, Edge) and operate seamlessly on both desktop and mobile devices (related to all SRs)

This concludes the requirement analysis, which will be looked at during the prototyping time of the 3rd research question.

3. Research Question 1

How can nursing home data systems ensures the integrity of their data (complete, accurate and consistent) through the development of a protocol?

3.1. Reason

The reason for the research question is to find out what data integrity means, are there any frameworks that could help in making the protocol that ensures data integrity during the life cycle.

3.2. Method

Figure 2 Dot framework methods used in this question

3.3. Result/product

Literature study:

What is Data quality ?

Data quality is basically how good or accurate a set of data is. It's important for businesses because it helps in making better decisions. Bad data can cause problems like treating the wrong patient or missing some patient's information. Fixing these problems can cost a lot of money (Sheldon, 2024). Good data quality means data is as follow:

1. Accuracy: Data accurately represents the entities or events it describes and comes from verifiable and trustworthy sources (The data is telling the truth.).
2. Completeness: Data includes all expected values and types, along with any accompanying data (All the mandatory fields such as BSNs are expected to be complete).
3. Consistency: Data is uniform across systems and datasets, with no conflicts between the same data values (BSN on system 1 should belong to the same patient in system
   2).
4. Validity: Data conforms to defined business rules and parameters, ensuring proper structure and content (Valid data so a person can't be born after his death).
5. Timeliness: Data is up-to-date and available when needed (Data is refreshed frequently).
6. Uniqueness: Data does not contain duplicate records within a dataset, and each record can be uniquely identified (No 2 patients should have the same number).

Businesses/organizations use automated checks and manual reviews to make sure their data meets these standards. Automated checks use tools or written scripts to detect errors and inconsistencies in data, while manual reviews involve human intervention to analyze data for accuracy and correctness. By identifying and resolving data errors, organizations can enhance the reliability and trustworthiness of the data, preceding to an developed decision-making processes and reliable business results (AI, 2022). This is done across all phases of the data life cycle, as it can be seen below.

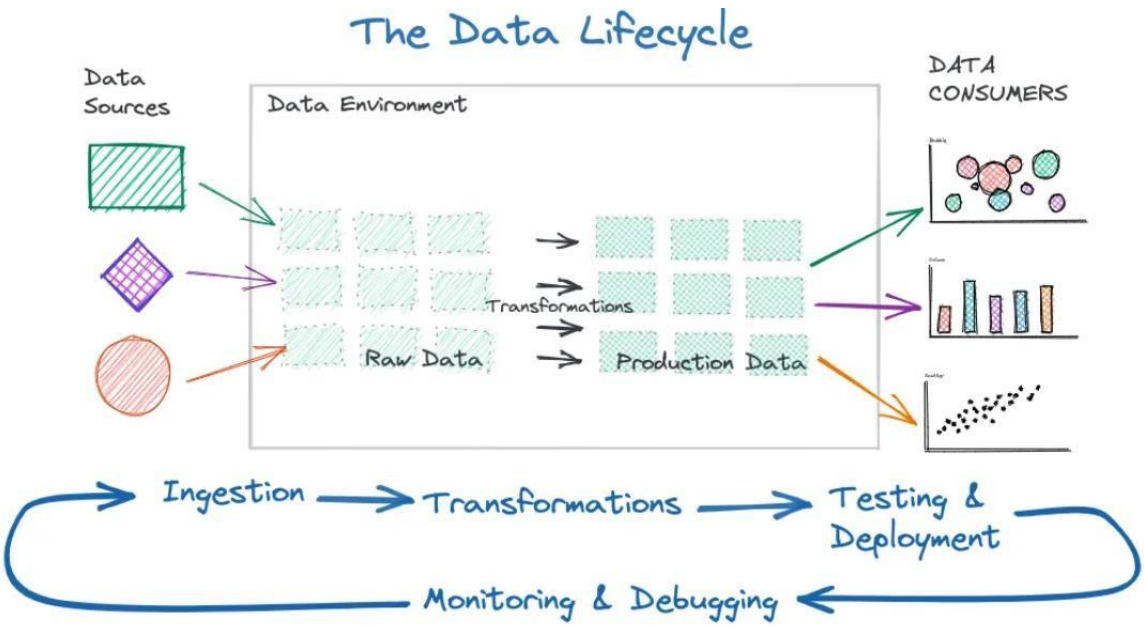


Figure 3 Data lifecycle

Data integrity is essential to overall data quality, embracing factors such as accuracy, completeness, consistency, and validity (LakeFS, 2024). With an understanding of data quality established, let's explore various data governance frameworks that can support the development of effective protocols.

What is data governance exactly ?

Data governance is a vital aspect of organizational management, essential for effectively managing and organizing the amounts of information produced by businesses. It involves establishing procedures, policies, and guidelines to guarantee the integrity of the data (ICTinformatiecentrum, 2023). With the exponential growth of data in recent years, the need for an effective data governance has become increasingly apparent, leading to the development of numerous data governance frameworks.

Data Governance Frameworks:

Several data governance frameworks offer organized methodologies to administer and protect sensitive data within organizations. These frameworks integrate guidelines for data management processes, rules, standards, and guidelines. Among the widely used frameworks are:

1. DGPO's Data Governance Best Practices Framework

2. DAMA Data Governance Framework
3. COBIT 5 Data Governance Framework
4. Open Data Institute (ODI) Data Governance Framework
5. Data Governance Council (DGC) Framework
6. Open Data Governance Framework (ODGF)
7. Data Management Body of Knowledge (DMBOK) Governance Framework

Each framework offers distinct strengths and may accommodate to specific industry requirements. For instance, some frameworks focus on improving data quality, while others emphasize regulatory compliance or open data governance. Organizations must assess their unique needs and priorities to select the most suitable framework.

Pros and Cons of these frameworks:

The frameworks vary in their pros and cons. For example, the DGPO framework provides comprehensive guidance across multiple categories but may require membership for full access. On the other hand, the DAMA framework offers a clear structure for data governance but might be perceived as too prescriptive/ rigid by some organizations (Incept Data Solutions, Inc. , 2023). The source that is just mentioned in the previous sentence, contains all the pros and cons of these frameworks.

Conclusion:

Choosing the right data governance framework is crucial for ensuring consistent and effective data management practices aligned with organizational goals. Factors to consider include organizational size, complexity of data management needs, existing processes, and available resources. Ultimately, the selected framework should facilitate accurate data-driven decisionmaking and support business success. In addition to that, due to the scope of the project, we will look at the data quality aspect of the governance within the DMBoK framework and will not be choosing a framework for the organization.

What is DMBOK?

DMBoK, which stands for Data Management Body of Knowledge, is a comprehensive framework developed by the Data Management Association (DAMA International) that provides guidelines and best practices for managing data effectively within

organizations. It encompasses various components and disciplines of data management and serves as a reference guide for data management professionals (Ganjhu, 2023).

The DMBoK framework includes several components, each focusing on different aspects of data management:

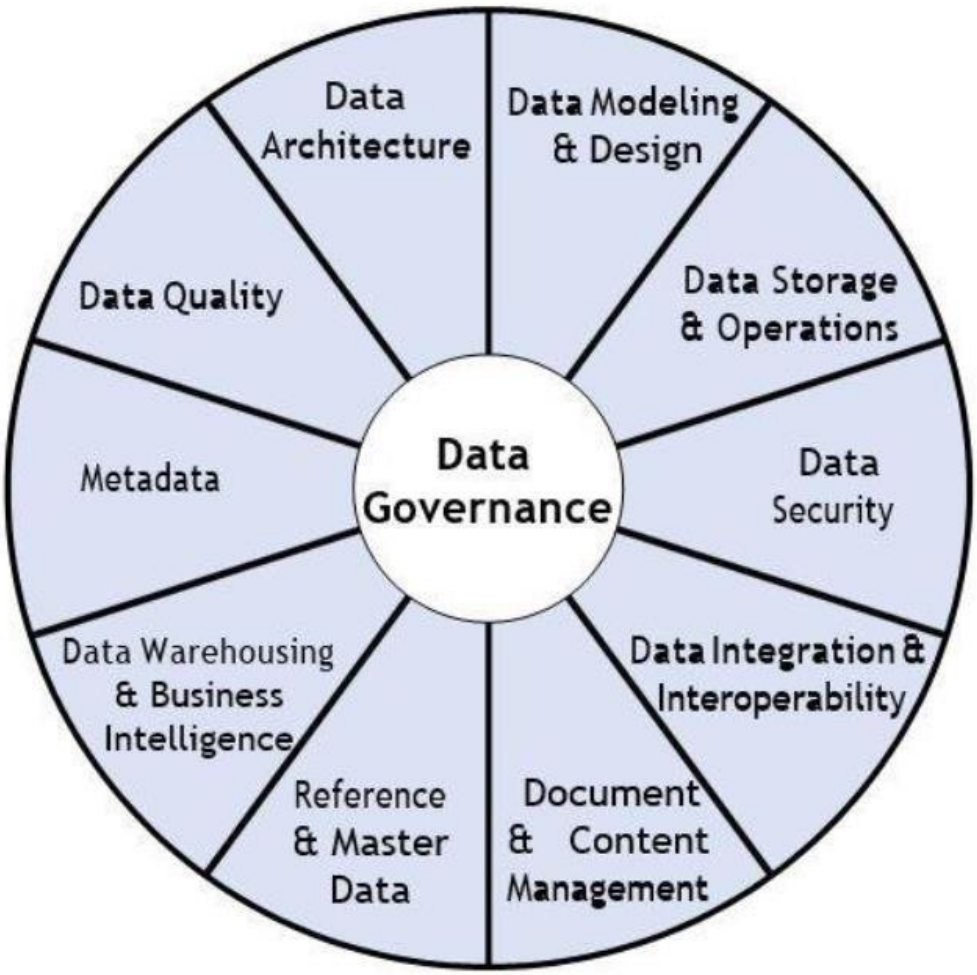


Figure 4 DMBOK Data governance diagram

the DMBoK framework offers several benefits and uses for organizations:

- **Comprehensive Guidance:** DMBoK provides comprehensive guidance and best practices for managing data across various disciplines, helping organizations establish effective data management strategies and processes.
- **Standardization:** By adopting DMBoK, organizations can standardize their data management practices and terminology, promoting consistency and alignment across different departments and projects.

- **Professional Development:** DMBoK serves as a valuable resource for data management professionals seeking to enhance their knowledge and skills, providing a structured framework for training, certification, and career development.
- **Adaptability:** DMBoK is designed to be flexible and adaptable to different organizational contexts and industries, allowing organizations to tailor its components and practices to meet their specific needs and requirements.

Compared to other frameworks, such as DGPO or COBIT 5, DMBOK offers a more holistic and comprehensive approach to data management, containing several aspects such as data quality, data governance, metadata management, data architecture, and more (Incept Data Solutions, Inc. , 2023), as it can be seen below.

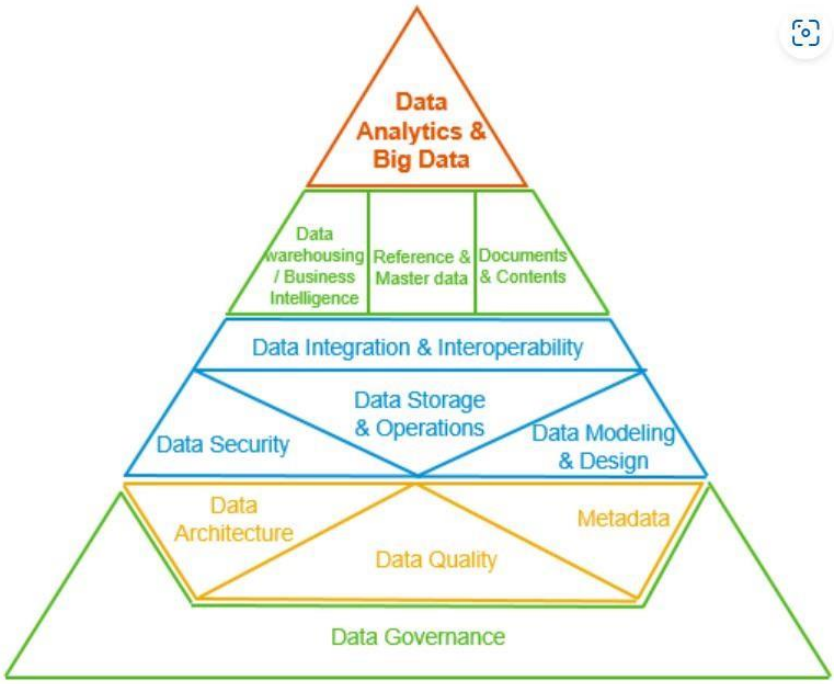


Figure 5 DMBOK Pyramid

Peter Aiken created the DMBOK pyramid, which outlines the relationship between the functional areas established by the DAMA Wheel in the figure above. It shows how data governance is the base of the pyramid then comes the orange layer. This makes the top layer the most important layer for the business (Ma, 2021). However, DMBOK provides detailed guidance and best practices for each aspect of data management, making it a valuable resource for organizations. Furthermore, DMBOK is widely recognized and used across industries, making it a reputable and trusted framework for data management and it is used in RABO Bank (Rabobank, n.d.).

Now that we know about the frameworks and DMBOK, let us look at the methodologies of ISO, and What is ISO 8000?

ISO 8000 is a set of rules created by a global organization called ISO. These rules are all about making sure that data (like information and numbers) used by companies is good quality. (ISO 8000, 2022) They give instructions and tips to companies on how to check if their data is accurate, complete, and consistent. The main goals of ISO 8000 are:

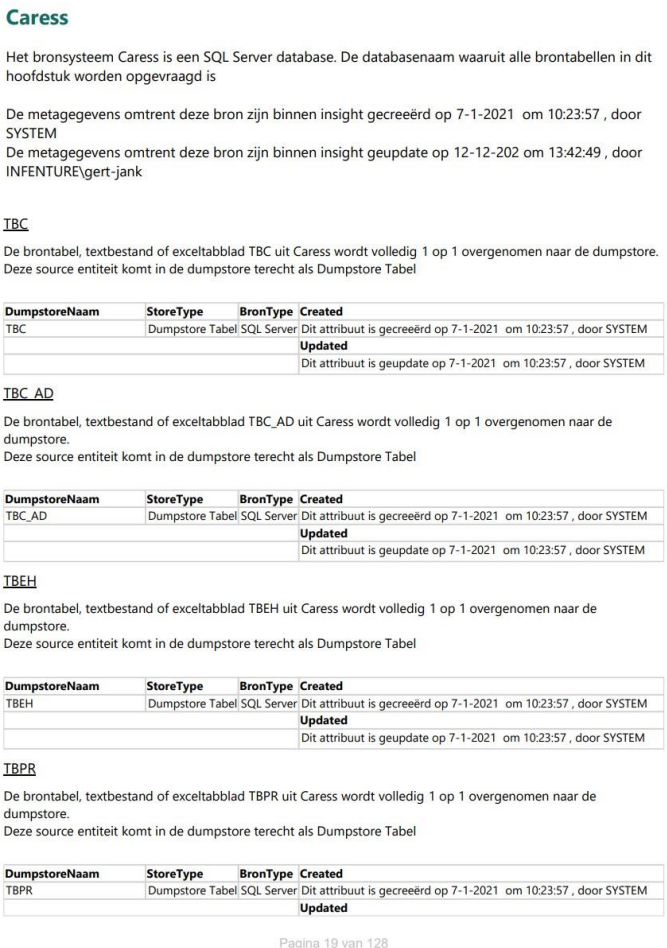
1. Making things the same: ISO 8000 helps companies make sure everyone agrees on what data means and how it's organized. This makes it easier for different systems and people to understand and work with the data.
2. Making data better: By following ISO 8000, companies can find and fix problems with their data, like mistakes or missing pieces. This leads to better quality data that helps with making decisions and doing business.
3. Sharing data easily: ISO 8000 gives guidelines on how to share data with others, like partners or customers, while keeping it accurate and trustworthy.
4. Working faster: ISO 8000 helps companies do things quicker and with fewer mistakes by setting up standard ways to manage data quality.
5. Avoiding problems: Bad data can cause big problems and costs for companies. ISO 8000 helps prevent these issues by making sure data is reliable and useful.

In short, ISO 8000 helps companies handle their data well, which leads to better decisions, smoother operations, and less risk in different industries and fields (Benson, ISO 8000: A new International Standard for Data quality, by Peter Benson , 2020).

Doc analysis:

During the first 2 weeks of the project, some documents were told are important to be read for the project. Below is the analysis of the documents.

1. First document- Align dumpstore documentatie: This document outlines the sources and source tables used in the data warehouse of AxionContinu. It describes the types of sources, such as Excel, Access, SQL Server, etc., and the tables or sheets within these sources from which data is imported or otherwise utilized in the Dumpstores. For each source table or sheet, it details whether it is used in its entirety or if there are additional processing queries applied, which are also provided if applicable. Additionally, it contain information on filters applied to source tables for Dumpstore purposes. The document categorizes source data into four types: Dumpstore Table, View in Dumpstore, Procedure or other source, and Temporary table. It also includes timestamps for when data extraction from a source and its tables was initiated and when it was last updated, which can be used in the quality check of timeliness. Below is an example of this document.



Caress

Het bronsysteem Caress is een SQL Server database. De databasenaam waaruit alle brontabellen in dit hoofdstuk worden opgevraagd is

De metagegevens omtrent deze bron zijn binnen insight gecreeërd op 7-1-2021 om 10:23:57 , door SYSTEM
De metagegevens omtrent deze bron zijn binnen insight geupdate op 12-12-202 om 13:42:49 , door INFENTURE\gert-jank

TBC

De brontabel, textbestand of exceltabblad TBC uit Caress wordt volledig 1 op 1 overgenomen naar de dumpstore.
Deze source entiteit komt in de dumpstore terecht als Dumpstore Tabel

DumpstoreNaam	StoreType	BronType	Created
TBC	Dumpstore Tabel	SQL Server	Dit attribuut is gecreeërd op 7-1-2021 om 10:23:57 , door SYSTEM
			Updated
			Dit attribuut is geupdate op 7-1-2021 om 10:23:57 , door SYSTEM

TBC_AD

De brontabel, textbestand of exceltabblad TBC_AD uit Caress wordt volledig 1 op 1 overgenomen naar de dumpstore.
Deze source entiteit komt in de dumpstore terecht als Dumpstore Tabel

DumpstoreNaam	StoreType	BronType	Created
TBC_AD	Dumpstore Tabel	SQL Server	Dit attribuut is gecreeërd op 7-1-2021 om 10:23:57 , door SYSTEM
			Updated
			Dit attribuut is geupdate op 7-1-2021 om 10:23:57 , door SYSTEM

TBEH

De brontabel, textbestand of exceltabblad TBEH uit Caress wordt volledig 1 op 1 overgenomen naar de dumpstore.
Deze source entiteit komt in de dumpstore terecht als Dumpstore Tabel

DumpstoreNaam	StoreType	BronType	Created
TBEH	Dumpstore Tabel	SQL Server	Dit attribuut is gecreeërd op 7-1-2021 om 10:23:57 , door SYSTEM
			Updated
			Dit attribuut is geupdate op 7-1-2021 om 10:23:57 , door SYSTEM

TBPR

De brontabel, textbestand of exceltabblad TBPR uit Caress wordt volledig 1 op 1 overgenomen naar de dumpstore.
Deze source entiteit komt in de dumpstore terecht als Dumpstore Tabel

DumpstoreNaam	StoreType	BronType	Created
TBPR	Dumpstore Tabel	SQL Server	Dit attribuut is gecreeërd op 7-1-2021 om 10:23:57 , door SYSTEM
			Updated

Pagina 19 van 128

Figure 6 Dumpstore doc

As it can be seen above, not much information can be deduced from this document.

2. Second document -Align infostore documentatie: This document details the entities loaded into the data warehouse of AxionContinu. Entities, categorized into dimensions, facts, or factless facts, contain attributes and are populated from one or multiple sources. It specifies the source, table or sheet, and column responsible for data within these entities. Entities are organized into Areas, indicating their grouping or loading context. Dimensions categorize facts and measures, facilitating business inquiry, often including time, product, employee, and cost center. Hierarchies within dimensions provide structured views, facilitating filtering, grouping, and labeling of measures. Facts represent measure values related to specific business processes, with each row denoting an event and its corresponding measure value, typically numeric and aggregatable. Factless facts are tables devoid of measure values, useful in scenarios such as tracking employee course attendance, offering flexibility in warehouse design and enabling simpler query execution. However, this document is meant for the Gold medallion phase of databricks and it is where the tables are expected to have quality. Below is an image of what can be seen in this document

Productie Care

Dimension Productie Care

Client

Basistabel	Doelkolom	Datatype	Brontabel	Bronkolom
TBC	Client_OKey	nvarchar (50)	TBC	I_C
		Dit attribuut is gecreeërd op 7-1-2021 om 10:23:57 , door SYSTEM Dit attribuut is geupdate op 7-1-2021 om 10:23:57 , door SYSTEM		
TBC	Client	nvarchar (100)	TBC	LONG_NM
		Dit attribuut is gecreeërd op 7-1-2021 om 10:23:57 , door SYSTEM Dit attribuut is geupdate op 7-1-2021 om 10:23:57 , door SYSTEM		
TBC	RegistratieNummer	nvarchar (10)	TBC	NR
		Dit attribuut is gecreeërd op 7-1-2021 om 10:23:57 , door SYSTEM Dit attribuut is geupdate op 7-1-2021 om 10:23:57 , door SYSTEM		
TBC	Voornaam	nvarchar (100)	TBC	NULLIF([TBC].[VR_NM],'')
		Dit attribuut is gecreeërd op 7-1-2021 om 10:23:57 , door SYSTEM Dit attribuut is geupdate op 7-1-2021 om 10:23:57 , door SYSTEM		
TBC	Roepnaam	nvarchar (100)	TBC	NULLIF(ISNULL([TBC].[call_nm],[TBC].[vr_nm]),'')
		Dit attribuut is gecreeërd op 7-1-2021 om 10:23:57 , door SYSTEM Dit attribuut is geupdate op 7-1-2021 om 10:23:57 , door SYSTEM		
TBC	Tussenvoegsel	nvarchar (10)	TBC	VRV
		Dit attribuut is gecreeërd op 7-1-2021 om 10:23:57 , door SYSTEM Dit attribuut is geupdate op 7-1-2021 om 10:23:57 , door SYSTEM		

Figure 7 Screenshot on infostore and the part that was helpful

As it can be seen above, this document only shows the size of the fields or the columns of the table, which helps in checking the consistency of the columns and their values.

3. Third document -Align Kubus Documentatie infent infostore: This document outlines the Kubus environment within AxionContinu. It specifies the measurement groups and values for each Kubus, along with how these values are linked to dimensions. It provides an overview of querying and filtering options for data analysis within the kubus environment. Additionally, it includes data types such as financial, patient data, etc., and details on folder structure, formulas, and code aggregations.

The first document describes the sources and tables used in AxionContinu's data warehouse, outlining how data is imported, processed, and filtered. It aids in maintaining data quality by ensuring accurate and relevant data is integrated into the warehouse. The second document outlines the entities within the data warehouse, categorizing them into dimensions, facts, and factless facts, along with their attributes and relationships. It helps ensure data quality by organizing and structuring data for efficient analysis. The third document details the kubus environment, specifying measurement groups, values, and their connections to dimensions. It assists in maintaining data quality by providing a structured framework for querying and filtering data during analysis, facilitating accurate insights. Collectively, these documents are found to be on a report level and not a source level. As the main goal of this project is to address the source level of the data. A screenshot of the documents can be found in appendices A-C, but they cannot be uploaded here or on canvas due to NDAs.

Best, good and bad practices:

Below is a standardized set of rules and processes including best, good and bad practices within the data quality dimension. Additionally, it includes actions on what needs to be done with an example of metrics:

1. **Accuracy:** data should mirror real-world situation from trustworthy source (ISO/TS 8000, 2012).
    - Best Practice: Regularly verify and validate data accuracy through automated consistency checks and manual reviews.
    - Action: Implement data validation rules and conduct periodic data audits.

- Example: Using checksums to verify data integrity in a database.
- Good Practice: Establish clear data entry procedures and provide training to personnel responsible for data entry.
- Action: Develop standardized data entry forms and provide training on accurate data input.
- Example: Requiring double-entry verification for critical data fields.
- Bad Practice: Relying solely on data without verifying its accuracy or integrity.
- Action: Ignoring data quality issues and continuing operations based on potentially inaccurate data.
- Example: Using outdated or incorrect customer address for visitation.

2. **Completeness:** where necessary value or fields are complete (Richman, 2023).
    - Best Practice: Ensure all required data fields are consistently populated and adhere to predefined standards.
    - Action: Implement mandatory fields in data entry forms and perform regular checks for missing data.
    - Example: Requiring all customer records to include both a name and contact information.
    - Good Practice: Establish data quality rules to flag incomplete records for review and completion.
    - Action: Implement automated alerts for incomplete records and provide tools for users to fill in missing information.

- Example: Notifying the business when a customer record lacks essential billing details or BSN number.
- Bad Practice: Allowing incomplete records to remain unresolved, leading to inaccurate reporting and decision-making.
- Action: Ignoring missing data fields and proceeding with data analysis or reporting.
- Example: Generating sales reports without complete customer information, leading to inaccurate revenue calculations.

3. **Consistency:** where same data in different storage aligns (Benson, ISO 8000 the International Standard for Data Quality , 2008).
    - Best Practice: Maintain consistent data formats, values, and definitions across systems and processes.
    - Action: Implement data standardization techniques and enforce data governance policies.
    - Example: Using standardized codes for product categories across all databases and applications. Mapping is also useful, to map the columns of the source with the lakehouse to ensure consistency.
    - Good Practice: Regularly resolve data inconsistencies between different systems to ensure consistency.
    - Action: Conduct data reconciliation processes and resolve discrepancies promptly.
    - Example: Comparing inventory levels recorded in the lakehouse system with those in the source system.
    - Bad Practice: Allowing inconsistencies between data sources to persist, leading to confusion and errors in analysis.
    - Action: Ignoring discrepancies between systems and proceeding with data integration or analysis.
    - Example: Combining sales data from two different systems without reconciling differences, resulting in inaccurate sales reports.
4. **Timeliness**: Data constantly available and up to date when required.

-
- Best Practice: Establish processes to ensure timely capture, processing, and dissemination of data.

  Action: Implement automated data capture mechanisms (timestamp) and set clear deadlines for data updates.

- Example: Using real-time timestamp to capture refreshed data for immediate analysis. If data timestamp =! Today, then timeliness needs to be checked.
- Good Practice: Monitor data arrival times and implement alerts for delays to ensure prompt action.
- Action: Set up automated notifications for late data arrivals and investigate reasons for delays.
- Example: Notifying team members when expected data uploads are overdue.
- Bad Practice: Allowing delays in data updates or reporting, resulting in outdated or irrelevant information.
- Action: Ignoring data update schedules and failing to address delays in data processing.
- Example: Using financial data from the previous month for current decision-making without verifying if more recent data is available.

5. **Validity:** Data aligns with acceptable formats and ranges (Richman, 2023)**.**
    - Best Practice: Establish data validation procedures to ensure compliance with predefined rules.
    - Action: Validate data against predefined criteria to ensure it meets quality standards.
    - Example: Verifying that all email addresses collected from web forms contain the "@" symbol and a domain name.
    - Good Practice: Provide feedback to the business on data validation errors to facilitate correction.
    - Action: Display informative error messages indicating why data validation failed and how to rectify it.

-
- Example: Notifying users when a BSN number is exceeding the 9 characters.
- Bad Practice: Allowing invalid data to be entered into the system without validation.

  Action: Allowing users to submit forms with invalid data without any indication of errors.
- Example: Accepting a birthdate of "99/99/9999" without validation, leading to inaccurate age calculations.

6. **Uniqueness:** All values should be unique with an ID (Anwar, 2024) **.**
    - Best Practice: Enforce uniqueness constraints on key data fields to prevent duplicates.
    - Action: Implement unique indexes or constraints on fields such as customer ID.
    - Example: Ensuring that no two customers in a database share the same email address.
    - Good Practice: Regularly deduplicate data to maintain a single source of truth.
    - Action: Identify and merge duplicate records based on common identifiers.
    - Example: Combining multiple entries for the same product into a single master column .
    - Bad Practice: Allowing duplicate records to proliferate without reconciliation.
    - Action: Allowing identical entries to be created in the system without detection or resolution.
    - Example: Creating multiple customer profiles for the same individual due to variations in spelling or formatting of names.

These are the protocol rules that every data engineer should keep in mind when working within any part of the data life cycle, to ensure the quality and integrity of the data (IBM documentation. , 2021).

Interview:

3 interviews were conducted for the purpose of this research question.

1. The first interview was with a data scientist. The purpose of the interview is to understand more about data quality and to get a perspective of a data scientist into the matter. During the interview, the most important things that we talked about were, how some machine learning techniques could be used to analysis the data quality and she elaborated more on how important it is to understand the data before doing anything. Interview can be found in appendices D.
2. The second interview was with a data engineer who was working by rabobank. The purpose of the interview was to get information from a fellow data engineer on how to approach this problem and to get to know how it was done in Rabobank and to validate the best practices. This interview had an impact on the best practices and the protocol made prior to this method, as we talked about her best and bad experiences in handling data quality within Rabo bank. The interview can be found in appendices E.
3. The Third interview was with an IT teacher. The purpose of that interview was to collect information on how he approached a similar situation and to validate the literature study. Additionally, a widespread discussion on various aspects of data management, including integrity, uniqueness, timestamping, hashing, primary keys, and business definitions. Key insights were shared regarding the importance of understanding business requirements, ensuring data integrity and uniqueness, and the need for collaboration between IT and business stakeholders. The interview can be found in appendices F.

Guideline conformity analysis:

To ensure the credibility of the quality of the data, some guidelines and best practices are selected from ISO 8000 standards and DMBOK for the data quality dimensions (Richman, 2023). Per dimension, a testcase has been made with a hypothesized working result, to verify the compliance of the guidelines to the protocol (Linkedin, 2024).

1. **Accuracy:**
    - **Objective:** To verify that data accuracy is maintained according to defined practices.
    - **Test Case:** Select a sample of data records and verify accuracy through automated script checks (i.e. # of nulls, duplicates en format) and manual reviews.
    - **Expected Results:** The selected data records pass automated script checks without errors, and manual reviews confirm the accuracy of the data against known standards.
2. **Completeness:**
    - **Objective:** To ensure all required data fields are consistently populated (LakeFS., 2024).
    - **Test Case:** Randomly select records and check if all mandatory fields are populated.
    - **Expected Results:** All selected records have populated mandatory fields according to predefined standards.
3. **Consistency:**
    - **Objective:** To ensure consistent data formats, values, and definitions across systems.
    - **Test Case:** Compare data values across different systems to identify inconsistencies.
    - **Expected Results:** Data values match consistently across all systems without significant variations.
4. **Timeliness:**
    - **Objective:** To verify that data updates are captured and disseminated in a timely manner.
    - **Test Case:** Monitor data arrival times and compare them against predefined deadlines.
    - **Expected Results:** Data arrives within the expected time frame, and any delays are promptly identified and addressed.
5. **Validity:**
    - **Objective:** To ensure that data conforms to predefined validation rules.

- **Test Case:** Input invalid data and verify that it is rejected by the system.
- **Expected Results:** The system rejects invalid data and provides informative error messages indicating the reason for rejection.

6. **Uniqueness:**
    - **Objective:** To prevent duplicate records in key data fields.
    - **Test Case:** Attempt to input duplicate records and verify that they are rejected by the system.
    - **Expected Results:** The system prevents the creation of duplicate records and provides feedback indicating that the record already exists.

Each of these test cases aims to verify compliance with the specified data quality dimensions (HU, 2023) and ensures that data adheres to the defined best practices, good practices, and avoids bad practices as outlined in the ISO 8000 guidelines.

3.4. Conclusion/ Evaluation

Literature study:

The literature study provided a comprehensive overview of data quality, data governance frameworks, the Data Management Body of Knowledge (DMBoK), and ISO 8000 methodologies. It emphasizes the importance of ensuring data accuracy, completeness, consistency, validity, timeliness, and uniqueness for effective decision-making in businesses. The study highlights various data governance frameworks available, each with its pros and cons, and stresses the need for organizations to select a framework aligned with their specific needs and priorities. Additionally, the DMBoK framework is presented as a holistic guide for managing data effectively within organizations. Moreover, ISO 8000 is introduced as a set of rules aimed at improving data quality, standardization, efficiency, and risk mitigation across industries. Overall, the literature study underscores the significance of data management practices supported by appropriate frameworks and methodologies, which will be crucial for informing and guiding the project towards creating the protocols.

Document analysis:

The document analysis did not add much into this question, as the documents provided were looking at the bigger picture and not to the source of the errors. Except for the Infostore and Kubus documents, as infostore provided information about some of the characteristics of the columns (i.e, length of column) and Kubus provided information about the types of data that exits within mijncaress (i.e., patients data, employee and etc..)

Best, good and bad practices:

Based on the literature study done and the interviews a protocol of best practices was made, as this should help any engineer in ensuring data quality. As This was made while rereading the literature again and got enhanced further via the interviews.

Interview:

The interviews underscores the critical role of both technical and business considerations in effective data management to ensure data integrity and its alignment with organizational goals. In addition to that, based on the interviews the protocol and the test cases were made.

Guideline conformity analysis:

The conclusion of the guideline conformity analysis emphasizes the need to follow established data quality standards, such as ISO 8000 and DMBOK, to maintain credible data for the project. Test cases for accuracy, completeness, consistency, timeliness, validity, and uniqueness are outlined to verify devotion to these standards. By implementing these tests, the project can ensure reliable and trustworthy data, reducing the risk of inaccuracies and supporting informed decision-making. Based on the experiences within the interview, testcases were developed per data quality dimension.

4. Research Question 2

What metrics or code of validation measures should be taken into account for ensuring data integrity, and can they be tailored for business requirements?

4.1. Reason

The reason for this question is to research which measures can be used to validate and analyze the current data and to decide which dimension fits with what column.

4.2. Method

Figure 8 Dot framework methods used for this question

4.3. Result/product

Literature study:

The Advised protocol was created from research findings, identifying useful metrics for analyzing current data and the business requirements. Below is a screenshot of these metrics which can be found in the protocol. To view the complete protocol, please refer to Appendices G.

Based on the research done above and the guidance meetings with the stakeholder, the business rules were made and the metrics are also chosen based on them. The following metrics below are taken from 3 different sources and they are (Mssaperla., 2024), (Aliaga, 2023) and (Databricks, 2024).

A. Completeness:
   i. Business:
      - Check required fields: Specify which columns are important for decision making.
      - Analyze completeness % on the columns individually, then per couple of columns combined.
      - Cross-referencing: maybe cross reference afas with miincaress.

   ii. Metrics:
      - Percentage of Missing Values: (Number of missing values / Total number of values) * 100
      - Completeness Score: (Number of filled fields / Total fields) * 100
      - Percentage of Records with Complete Information: (Number of records with no missing values / Total number of records) * 100
      - Timeliness of Data Entry: Measure the time taken to fill missing data.
      - Percentage of Mandatory Fields Filled: (Number of mandatory fields filled / Total number of mandatory fields) * 100

Figure 9 Part of the protocol made

Ethical check:

The ethical check aimed to align the company's norms with the data being analyzed, which primarily consists of patient information such as date of birth, name, and addresses. The assessment of data integrity dimensions involved collaboration with a colleague to determine suitable fits for each column. For instance, given the ethical consideration that individuals can change genders, accuracy was not applicable to the gender column. Subsequently, an Excel sheet was prepared as outlined below:

Tabelnaam	Dimensions	KolomNaam	Voorbeeld Data	Data betekenis	Mogelijke Filters	Mogelijke kwaliteitscontroles
CoreTables						
Tbc	unqiue en consistant	IC		BK		Wisselende waarden
	complete en consistant	LongNm		Naam met voorletters	Naar UCASE of upper/down case	?
	uniqnue en consitant	Nr		patient number?		?
	constiancy, completeness	VrNm		Voornamen	Naar UCASE of upper/down case	Kan 1 of 2 of meer namen bevatten
	constiancy	CallNm		Roepnaam	Naar UCASE of upper/down case	ca
	consitancy	Vrv		Voor voegsel	Naar UCASE of upper/down case	c
	consistancy	AchtNm		Achter naam	Naar UCASE of upper/down case	?
	niks	Geslacht		Geslacht	Upper case	M of V --> in infostore Man/Vrouw
	accuracy, completeness	GebDat		Geboorte datum		Toekomst?
	accuracy	BirthPlace		Geboorte plaats	Naar UCASE of upper/down case	?
	accuracy	OvlDat		Overleidingsdatum		Na GebDat?, Toekomst?
	niks	Iusr		FK		Overal Null?
	check It out	Icountry		FK		?
	niks	Ilang		FK		Overal Null?
Tbc_Ad	unqinuess	ICAd		BK		?
	accuracy and complentess	Straat		Straat naam	Naar UCASE of upper/down case	Bevat ".","?" en "1e/2e straatnamen" enz.
	accuracy	Hnr		Huisnummer		Huisnummers met 0 ervoor? Int van maken?
	accuray	HnrToe		Huisnummer toevoeging	Naar UCASE of upper/down case	?
	accuracy	Pc		Postcode	Upper case	Woon
	accuracy	Woonpl		Woonplaats	Naar UCASE of upper/down case	Bevat ".","?" en hoofdletters enz., Corrigieren op basis van postcode tabel? Omang met upercase/lowercase and speciale tekens
	check It out	ICountry		FK		Enige waarden die voorkomt of Null
	niks	IPcountry		FK		Enige waarden die voorkomt of Null
	unqinuess en constant	IC		FK naar Tbc		Wisselende waarden

Figure 10 Excel that shows the rules and the checks per column

For NDA reasons, some columns might be blurred, like the voorbeeld data in the figure above. Additionally, during the ethical check, it was realized that not all the dimensions of data integrity could be measured, as an access to a different source of the same data is mandatory to measure accuracy, uniqueness and completeness.

Data quality checks:

After the research and preparations were done, the time to apply the metrics has come. This is done in a databricks notebook, within AxionContinu's environment. Within the notebook, the data is loaded and is ran through the same pieces of code that checks the consistency of all the columns within the tables. Additionally, some inaccuracies were found as it can be seen below, the date of birth is after the date of death.

invalid_rows

	VR_NM	VRV	ACHT_NM	GESLACHT	GEB_DAT	BIRTH_PLACE	OVL_DAT	TEL1	BSN	SOFI_NR
21161		None		V	1946-05-07	Beni Badis	1921-09-27	(		

Figure 11 Example of errors found during analysis

Furthermore, a metric for the BSN was made to check the validity of the BSN (11proef) and consistency. ****** Add the 11 proof reference here. The notebook cannot be uploaded to canvas, but a screenshot of it can be found in the appendices H. However, based on the analysis there the excel sheet was further filled.

a betekenis	Mogelijke Filters	Mogelijke kwaliteitscontroles
		Wisselende waarden
m met voorletters	Naar UCASE of upper/down case	All rows starts with capital letter
ent number?		niks
rnamen	Naar UCASE of upper/down case	Kan 1 of 2 of meer namen bevatten & niet alle rijen begint met een capital letter
pnaam	Naar UCASE of upper/down case	niks
r voegsel	Naar UCASE of upper/down case	Most are Within range, however 1 line with 10 characters and looks wrong " aan dr wal-" included a last name
ter naam	Naar UCASE of upper/down case	All capitalized
lacht	Upper case	M of V --> in infostore Man/Vrouw no 3rd value was found
oorte datum		first available date is january 1900 and last available date is novemeber 2022. 1 row found where DOB > DOD and there are around 16 rows with NA dob
oorte plaats	Naar UCASE of upper/down case	Not all rows are capitalized sometimes Utrecht sometimes U. 21668 Rows with accents and 440 without.
rleidingsdatum		There are dates in the past 1920s latest date available march 2024, 1 row where dod was before dob.
		Overal Null?
		?
		Overal Null?
		So many inconsitancies 3130- number or 3130 " space"then number or numbers starting with 0000000 or 3130thennumber no pace or +31 or 0031 or starts with 31 and no 0
		This is moved to BSN
		Alll are valid and passed the test

Figure 12 Excel sheet that contains the patterns found the rules and what to check and how

Interview:

An interview was done with Linda who is a data engineer at twenty next and used to work with rabo bank. The purpose of the interview was to validate the advice on the ETL process and the metrics chosen in measuring the consistency of the data. The interview can be found in Appendices I. However the interview can be summarized into the following :

- Good metrics and rules chosen for the ETL processes and asked for this to be shared with the others.
- In this project, consistency is the only valid dimension to look at, as to check completeness, accuracy and uniqueness, we will either need the main data before bronze medallion or a different data set for cross reference.

4.4. Conclusion/Evaluation

Literature study

Based on the research done on valuable metrics, keeping DMBOK and iso8000 in mind, an advised protocol document was made. The document includes the definitions of data integrity from the business perspective, Roles and Responsibilities, training and education

and translation of the business requirements and metrics. The document can be found in Appendices G

Ethical check

During the ethical check, it was realized that not all the dimensions of data integrity can be measured within these tables. As these are patients data and most of the errors are caused due to human input error (i.e., input of “? ” or space instead of null or Nan). Therefore an excel sheet was made to document the findings and what can be checked within consistency. A screenshot can be found in figure 7 due to NDA the document cannot be uploaded to canvas.

Data quality check

This method is usually used for machine learning, however in this case it is a requirement from the business. So the consistency of the current data was tested per column per row. Some inaccuracies were found as mentioned in chapter 3.3, however more were noticed during the Exploratory data analysis.

Interview

The interview got to confirm and review the protocols and the metrics that were used to measure the consistency of the data. Additionally, it confirmed that not all dimensions can be used to measure all the columns of these 2 tables. Furthermore, the protocols and the notebook should be shared with the engineers as they want to use it, as this is not written anywhere.

5. Research Question 3

How can an integration into the IT landscape be incorporated, while measuring the dimension of data integrity and employing suitable visualization in a monitoring dashboard ?

5.1. Reason

The purpose of this question is to research how to connect the data of AxionContinu to Power BI, and visuals that can be used for monitoring data within the prototype.

5.2. Method

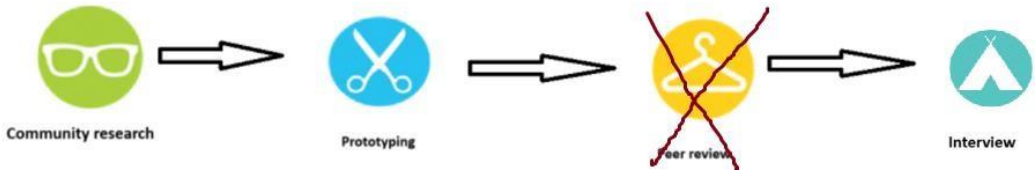


Figure 13 Dot framework methods used for this question.

The Method has changed compared to the project plan, as peer review should be used to collect feedback and review the product by an external expert and not within the company. Therefore, Interview is a better choice.

5.3. Result/product

Community research

The first part of community research was about how to get the data into Power BI. It was discovered that databricks has its connector within Power BI, which can be used to get the data in, as it can be seen below.

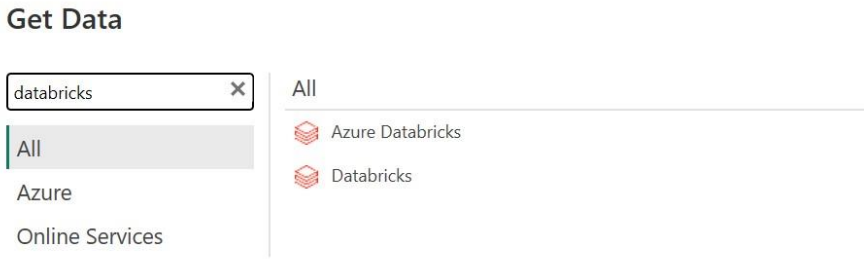


Figure 14 showing how to connect to databricks.

After clicking on the connector and signing in with the Username and password that was provided to me, I can navigate to the bronze warehouse and pick the tables that needs to be loaded (Databricks, 2024).

Additionally, before conducting community research on visualizations, Python Databricks codes were being translated into DAX Power BI codes with the assumption that they should also be visualized in Power BI, as it can be seen below.

```
DOD_Date_Statistics =
VAR MinDate = MIN('mijncaress_crsadmin_tbc'[OVL_DAT])
VAR MaxDate = MAX(mijncaress_crsadmin_tbc[OVL_DAT])
VAR NullCount = COUNTROWS(FILTER('mijncaress_crsadmin_tbc', ISBLANK('mijncaress_crsadmin_tbc'[OVL_DAT])))
RETURN
    "Min Date: " & FORMAT(MinDate, "Short Date") &
    ", Max Date: " & FORMAT(MaxDate, "Short Date") &
    ", Null Count: " & NullCount
```

Figure 15 showing the manual code that was made manually.

Little was known about the built-in function that does the EDA (Data profiling) for you in Power BI. Based on the community resarch posted by radacad (Rad, 2019), it was discovered that in a new blank query you can profile the data and then display it in a table thorugh a simple code, as it can be seen below.

```
= Table.Profile(mijncaress_crsadmin_tbc)
```

Figure 16 Showing the code found after the research.

This code simplifies the translation of the coding languages, however not for the BSN 11-proef (which was hard to do in dax)nor the date checks, as these needs to be done manually.

Prototyping

For this method, a Power BI dashboard was made to visalize the error and the profiling of the data. This is one of the end products that will be delivered to the stakeholder but it's only a prototype and will not be pushed to production. However, after the commmunity resarach and the data being collected, the building of the dashboard was commenced as follow:

- The data model/semantic model with couple of tables was built as it can be seen below.

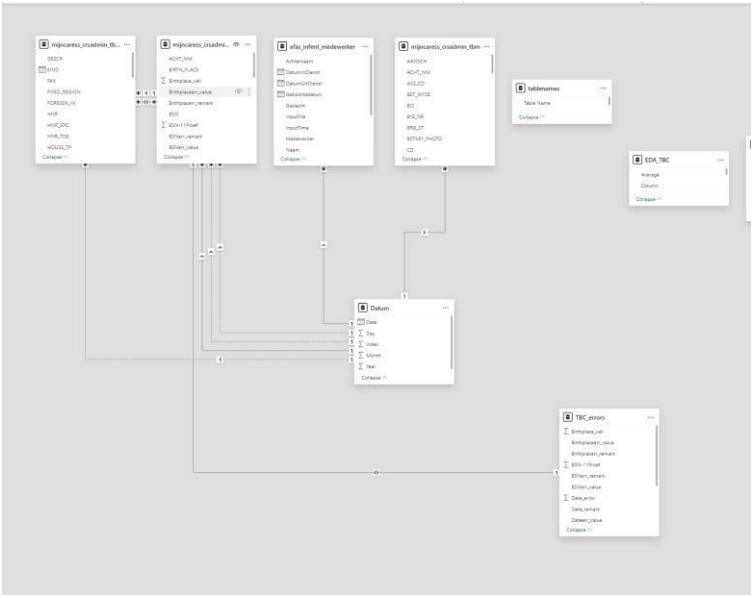


Figure 17 Semantic model

- The first template of the dashboard was also made with a home page and the information or the EDA that was manually made, as it can be seen below.

Column	Min	Max	Average	StandardDeviation	Count	NullCount	DistinctCount	Uppercase	Lowercase
ACHT_NM	Aa	Ünlü			33334	0	12275	TRUE	FALSE
BIRTH_PLACE		Üzengili			33334	777	2963	TRUE	FALSE
BSN	0	[redacted]			33334	582	32730	TRUE	FALSE
BSN-11Proef	0	1	[redacted]	[redacted]	33334	0	2	FALSE	FALSE
BSNerr_remark					33334	0		FALSE	FALSE
BSNerr_value					33334	582		FALSE	FALSE
Birthplace_vali	0	1	0,029789404	0,170008419430196	33334	0	2	FALSE	FALSE
Birthplaceerr_value					33334	777		FALSE	FALSE
Birthplacerr_remark					33334	777		FALSE	FALSE
Date_error	0	1	5,99988E-05	0,00774577304353794	33334	0	2	FALSE	FALSE
Date_remark					33334	0		FALSE	FALSE
Dateerr_value					33334	0		FALSE	FALSE

Figure 18 EDA made in Power BI

Figure 19 Home navigation page for the dashboard

- The BSN 11-proef was made in dax but with some errors. As in databricks when the BSN were validated, all the BSNs were valid, and some were empty.

  However, in DAX, I struggled with multiplying the indexes which resulted in false results, as it can be seen below.

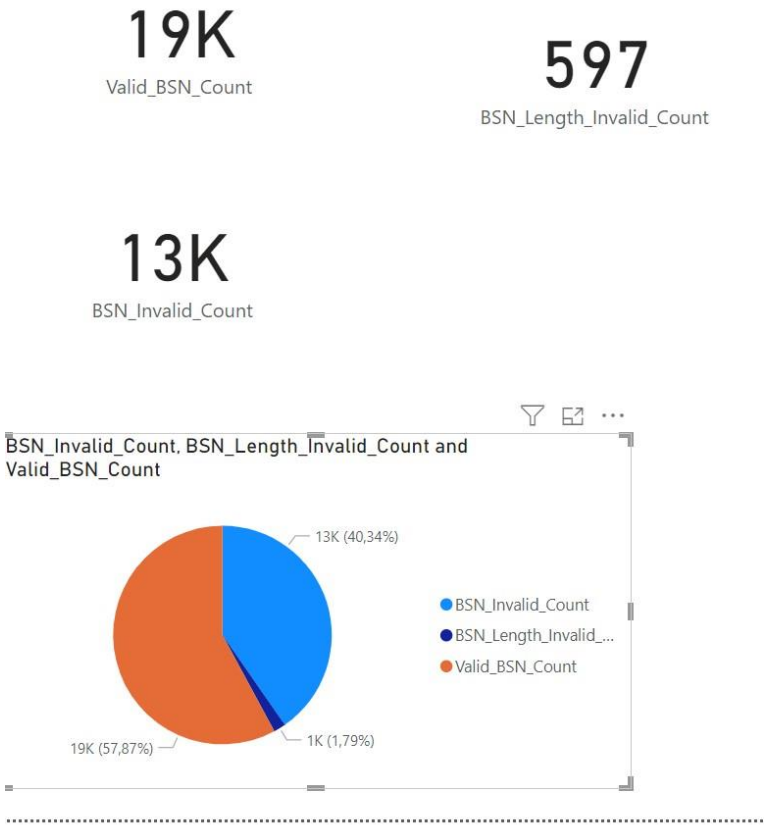


Figure 20 BSN invalid struggle

As it can be seen above, it says that there are 13k BSN numbers that are not valid. However, this was fixed with the help of my mentor, as he mentioned that I need to “divide and conquer” the code which worked.

Interview

After building the prototype, an interview was mode with the mentor/ boss Maikel, to validate the prototype and collect feedback for iteration. During the review, the dashboard was presented, but the boss expressed dissatisfaction, stating that it primarily showcased the exploratory data analysis (EDA) or data profiling rather than highlighting errors. Additionally, it was noted that I continued to focus on the analysis phase. I explained to the boss that Fontys' preference is to continuously iterate between analysis and prototyping. For instance, they don't share the belief that the analysis phase can be considered complete, as desired by the stakeholder. Nonetheless, feedback was also collected on how they would like the errors to be visualized with the requirements of the drill through, as it can be seen below.

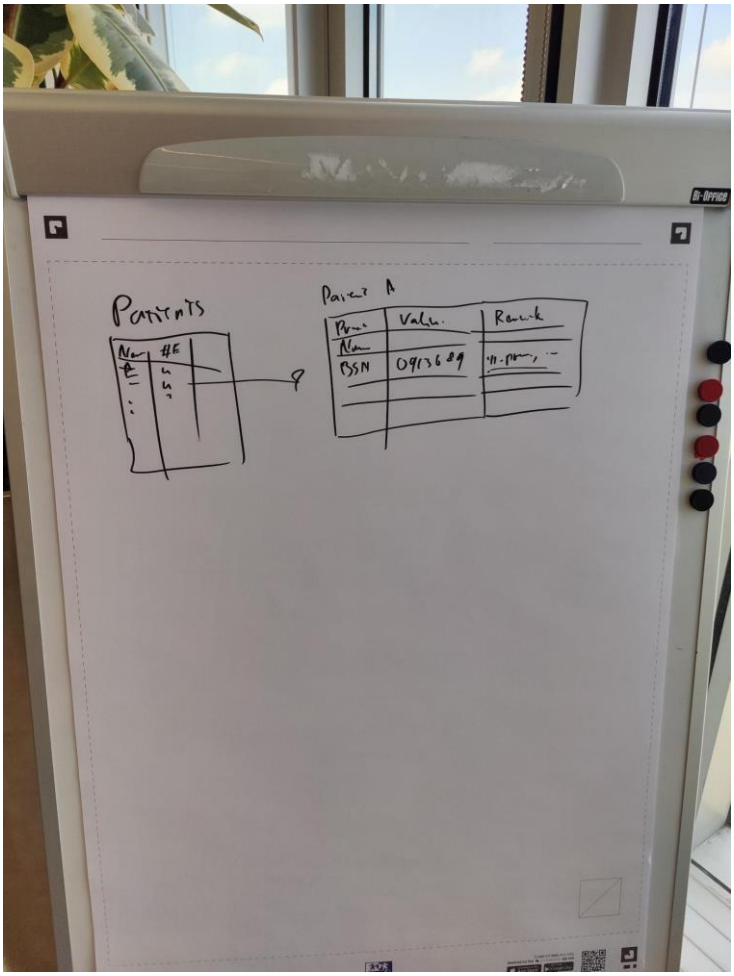


Figure 21 Stakeholder requirments in a diagram

The figure above shows the requirement of the stakeholder which can be summarized as follow:

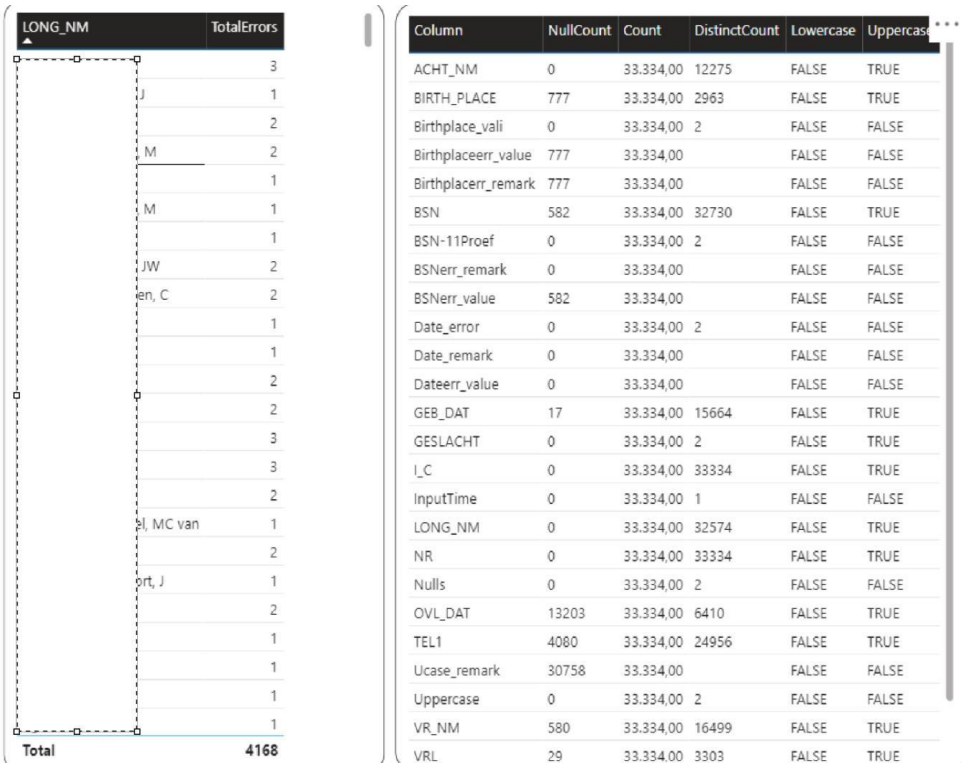
1. First page displays the patients and the amount of errors per row (includes BSN check, Date Check, Ucase checks and etc).
2. From the first page a drill through to the second page which shows the errors and their values in this format:

BSN Error	1231232134	BSN does not pass the 11 proef
Ucase error	hein-jan	The name didn't start with a capital letter.

3. Another visual that displays the error and the amount of patients that has that error and when you drill through you see all the patients who has that error in their record.

Prototype 2

Based on the review and feedback collected, the dashboard has improved to the wishes of the stakeholder, as it can be seen in the figure below.



LONG_NM	TotalErrors
	3
J	1
	2
M	2
	1
M	1
	1
JW	2
en, C	2
	1
	1
	2
	2
	3
	3
	2
l, MC van	1
	2
rt, J	1
	2
	1
	1
	1
	1
Total	4168

Column	NullCount	Count	DistinctCount	Lowercase	Uppercase
ACHT_NM	0	33.334,00	12275	FALSE	TRUE
BIRTH_PLACE	777	33.334,00	2963	FALSE	TRUE
Birthplace_vali	0	33.334,00	2	FALSE	FALSE
Birthplaceerr_value	777	33.334,00		FALSE	FALSE
Birthplacerr_remark	777	33.334,00		FALSE	FALSE
BSN	582	33.334,00	32730	FALSE	TRUE
BSN-11Proef	0	33.334,00	2	FALSE	FALSE
BSNerr_remark	0	33.334,00		FALSE	FALSE
BSNerr_value	582	33.334,00		FALSE	FALSE
Date_error	0	33.334,00	2	FALSE	FALSE
Date_remark	0	33.334,00		FALSE	FALSE
Dateerr_value	0	33.334,00		FALSE	FALSE
GEB_DAT	17	33.334,00	15664	FALSE	TRUE
GESLACHT	0	33.334,00	2	FALSE	TRUE
I_C	0	33.334,00	33334	FALSE	TRUE
InputTime	0	33.334,00	1	FALSE	FALSE
LONG_NM	0	33.334,00	32574	FALSE	TRUE
NR	0	33.334,00	33334	FALSE	TRUE
Nulls	0	33.334,00	2	FALSE	FALSE
OVL_DAT	13203	33.334,00	6410	FALSE	TRUE
TEL1	4080	33.334,00	24956	FALSE	TRUE
Ucase_remark	30758	33.334,00		FALSE	FALSE
Uppercase	0	33.334,00	2	FALSE	FALSE
VR_NM	580	33.334,00	16499	FALSE	TRUE
VRL	29	33.334,00	3303	FALSE	TRUE

Figure 22 Improved visuals based on requirements

In the figure above, on the left side table, the patients are displayed with the count of total errors per patient and once you drill through a patient it takes you to a different page as required.

LONG_NM	[redacted]
[redacted]	
Birthplaceerr_value	?
Birthplacerr_remark	Can contain the following : @, *, #, ?, !, &, `, |, \, /, (, ., $, _
BSNerr_value	
BSNerr_remark	Invalid length or failed 11 proef
Dateerr_value	DOD-No error
Date_remark	DOD-No remark
First Ucase_remark	Check Ucase of the following columns: 1. Long_NM, 2. Geslacht, 3. Voornaam en achternaam 4. Birthplace

Figure 23 Improved drill-through visual based on requirements

As it can be seen above, this row or this patient has BSN missing and has a “?” as his birthplace. Additionally, also based on the review, the third requirement is also made as it can be seen below.

BSN-11Proef	Count of LONG_NM
0	32737
1	597

Birthplace_vali	Count of LONG_NM
0	32341
1	993

Date_error	Count of LONG_NM
0	33332
1	2

Nulls	Count of LONG_NM
0	2541
1	30793

Figure 24 showing the amount of patients per error.

This visual is showing the count of patients per error. Furthermore, during the building of the dashboard, bookmarks were discovered and were used for the design of the dashboard. The figure below, shows the filters on the left side of the screen, these filter uses a bookmark to save space.

Figure 25 Filter button closed

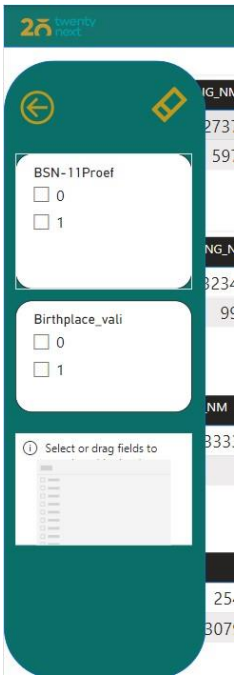


Figure 26 Filter page open

Additionally, a new feature was discovered while building the dashboard and that is showing the applied filter or filters in use without taking much space in the dashboard page, as it can be seen below.

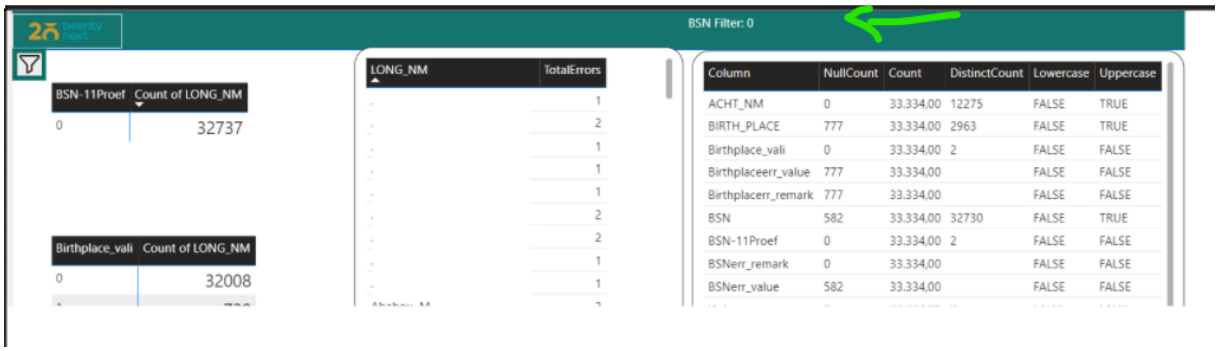


Figure 27 Applied filters without taking space

These filters remain empty when no filtering is applied but become visible when any filtering is activated. This Dashboard will go through a third iteration within the thesis document.

5.4. Conclusion

Community research

The first part of the research looked into how we can bring data into Power BI. We found that Power BI has a connector for Databricks, which makes it easy to import data. After signing in, we can choose which tables we want to load from the Databricks bronze warehouse. Before we started looking into different ways to show data visually, we were translating Python Databricks

code into Power BI's DAX code, thinking it could also be visualized in Power BI. Later, we learned about a feature in Power BI that can do data profiling for us, making it easier to understand our data. This feature helped simplify coding, but some checks, like BSN and date checks, still needed to be done manually.

Prototyping

The development process involved creating a Power BI dashboard to visualize errors and data profiling. While this prototype won't be pushed to production, it served as a crucial step in gathering stakeholder feedback. Key stages included building the data model, initial dashboard templates, and resolving issues with BSN validation. With mentor guidance, improvements were made, aligning the dashboard with stakeholder requirements. Notable features include patient error counts, drill-through functionality, and visualizing patient errors. The use of bookmarks optimized dashboard design, while dynamic filters enhanced user experience. Further iterations are planned to refine the dashboard's effectiveness.

Interview

After completing the prototype, an interview session was conducted with mentor/boss Maikel to validate the prototype and gather feedback for improvements. During the review, it was observed that the dashboard primarily focused on EDA rather than highlighting errors, leading to dissatisfaction from the boss. Additionally, concerns were raised about the continued emphasis on the analysis phase. However, it was explained that Fontys prefers an iterative approach between analysis and prototyping, contrary to the stakeholder's desire for a completed analysis phase. Feedback was also collected on visualizing errors and drill-through requirements. Stakeholder requirements were summarized, including displaying patient errors on the first page, drill-through to detailed error information on the second page, and a visual showing error distribution among patients.

6. Advice/recommendation

This chapter offers clear advice to help the organization solve problems and improve in areas like IT systems, compliance with laws, and strategic alignment. It provides practical strategies based on best practices to boost efficiency and ensure the

organization's goals are met. With tailored guidance, decision-makers can implement effective solutions for long-term success.

6.1. Previous situation

The previous situation of the project involved Axion Continu, a client of Twentynext, facing challenges in their data management process. Specifically, they encountered issues with data fragmentation, inconsistencies, and accessibility due to sourcing data and dashboards from four different companies: Infent, Congos, OGD, and MijnCaress. This fragmented approach led to missing data and delays, hampering Axion Continu's ability to access complete information necessary for their operations. To address these challenges, the proposed solution involved streamlining data sources from four to one platform. Progress had been made by collecting data from Software as a Service (SaaS) platforms and depositing it into an SQL lake. However, the transition between data layers, specifically from the bronze to silver layers within the Medallion Lakehouse architecture, posed significant risks to data reliability and accuracy.

The project aimed to enhance data quality and integrity within Axion Continu's operations by implementing validation measures, defining clear quality standards, and establishing quality assurance mechanisms. Additionally, the development of a dashboard for visualizing errors and identifying files with potential errors was planned to enable proactive management and decision-making.

Overall, the previous situation highlighted the critical need for improving data quality and streamlining data management processes to ensure accuracy, reliability, and accessibility of data for Axion Continu's operations.

6.2. Advised situation.

Compared to the previous situation, this chapter presents targeted strategies for further improvement within the company.

1. Integration with IT Landscape: For future references, databricks has its own dashboard and its own environment to validate the data, including what to do with errors and visualize the errors before ingesting the data into the bronze layer (Rachidi, 2024). This means that Power BI isn’t mandatory nor some validation measures (i.e., Date checks). However, A notebook could be introduced after the bronze layer and before the silver to validate some measures like the BSN-11 proef, as that is not achievable by databricks alone.
2. Focus on Error Highlighting: This involves ensuring that the dashboard prominently displays errors and their details, as per the stakeholder's requirements. The visualization should effectively communicate the nature and frequency of errors encountered in the data.

3. Iterative Development: Adopt an iterative development approach to refine the dashboard's effectiveness. Continuously gather feedback from stakeholders, and users to identify areas for improvement and implement necessary changes. This iterative process ensures that the dashboard evolves to meet changing requirements and user needs over time.
4. Enhanced User Experience: Pay close attention to optimizing the user experience of the dashboard. Utilize features such as bookmarks, dynamic filters, and intuitive navigation to enhance usability and accessibility. The dashboard should be user-friendly and easy to navigate for all stakeholders, including non-technical users.
5. Data Integrity Measures: Implement data integrity measures to maintain the accuracy and reliability of the data within the dashboard. This involves thorough validation of critical data elements such as BSN and date checks. Ensure that data validation processes are automated where possible and manual checks are performed accurately.
6. Stakeholder Alignment: Maintain alignment with stakeholder expectations and requirements throughout the project lifecycle. Regularly communicate with stakeholders to ensure that their needs are being met and that the dashboard delivers value in line with their goals and objectives.
7. Advice on Framework Selection: It is advised that the organization selects a specific framework that best fits its unique needs. This project and the accompanying advice have primarily focused on the data quality aspects of the DMBoK and ISO frameworks. While these two frameworks are recommended, it is ultimately the stakeholders who have the most comprehensive understanding of the organization's overall structure and requirements. Therefore, their insights should guide the final decision on framework selection.

Overall, the advised situation involves a holistic approach to the dashboard development, focusing on error highlighting, iterative improvement, integration with the IT landscape, data integrity, and stakeholder alignment. By following these recommendations, the project can deliver a high-quality data and dashboard that effectively meets the needs of its users and stakeholders.

7. References

Linkedin. (2024, 02 19). *How can you design data quality test cases?* Retrieved from Linkedin: https://www.linkedin.com/advice/0/how-can-you-design-data-qualitytest-cases-skills-data-quality

AI, C. (2022, 05 2). *Importance Of Data Quality In The Age of AI*. Retrieved from Medium: https://cetasai.medium.com/importance-of-data-quality-in-the-age-of-aif91ad2c49c7d

Anwar, M. (2024, 01 15). *5 Crucial Best practices for ensuring data quality in healthcare.* Retrieved from Astera. : https://www.astera.com/type/blog/managing-dataquality-in-healthcare/

Aziz, F. (2023, 07 11). *Requirement analysis process in software development.* Retrieved from InvoZone: https://invozone.com/blog/importance-of-requirement-analysisprocess-in-software-development/

Benson, P. (2008, 07 16). *ISO 8000 the International Standard for Data Quality* . Retrieved from The MIT 2008 Information Quality Industry Symposium: http://mitiq.mit.edu/IQIS/Documents/CDOIQS_200877/Papers/13_01_5A-1.pdf

Benson, P. (2020, 02 29). *ISO 8000: A new International Standard for Data quality, by Peter Benson* . Retrieved from Data Quality Pro. Data Quality Pro. : https://www.dataqualitypro.com/blog/iso-8000-new-international-standarddata-quality

Cox, A. (2017, 09 06). *Business Requirements vs Functional Requirements? Who Cares?* Retrieved from EN. Netmind. : https://netmind.net/en/business-vs-functionalrequirements-who-cares-en/

Cox, A. (2024, 05 13). *Software Requirements Specification (SRS):* . Retrieved from definition, example, how to write, & more. : https://www.inflectra.com/Ideas/Topic/Requirements-Definition.aspx

Databricks. (2024, 3 3). *Connect Power BI to databricks.* Retrieved from Databricks on AWS.: https://docs.databricks.com/en/partners/bi/power-bi.html

Doel, R. V. (2018, 01 12). *User Requirements*. Retrieved from https://perfval.com/userrequirements/

Ganjhu, P. K. (2023, 05 24). *The DAMA-DMBOK Functional Framework: A Comprehensive approach to effective data management.* Retrieved from Medium: https://pawankg.medium.com/the-dama-dmbok-functional-framework-acomprehensive-approach-to-effective-data-management-3de06af6

HU, K. (2023, 05 10). *How to set up data quality Tests*. Retrieved from Metaplane: https://www.metaplane.dev/blog/how-to-set-up-data-quality-tests

IBM documentation. . (2021, 03 05). *Data quality methodology*. Retrieved from IBM: https://www.ibm.com/docs/en/iis/9.1?topic=practices-data-qualitymethodology

ICTinformatiecentrum. (2023, 11 08). *Data governance framework*. Retrieved from ICTinformatiecentrum: https://www.ictinformatiecentrum.nl/datamanagement/data-governance/data-governance-framework

Incept Data Solutions, Inc. . (2023, 02 17). *WHICH DATA GOVERNANCE FRAMEWORK SHOULD YOU CHOOSE?* . Retrieved from LinkedIN: https://www.linkedin.com/pulse/which-data-governance-framework-shouldyou-

ISO 8000. (2022, 02). *Data quality*. Retrieved from Part 2: Vocabulary. : https://www.iso.org/obp/ui/#iso:std:iso:8000:-2:ed-5:v1:en

ISO/TS 8000. (2012, 11 3). *Data quality*. Retrieved from Part 311: Guidance for the application of product data quality for shape (PDQ-S).: https://www.iso.org/obp/ui/#iso:std:iso:ts:8000:-311:ed-1:v1:en

LakeFS. (2024, 03 11). *What is Data Quality?* Retrieved from Definition, Framework & Best Practices: https://lakefs.io/data-quality/

LakeFS. (2024, 03 11). *Data Quality testing:*. Retrieved from Ways to test data validity and accuracy. Git For Data - lakeFS. : https://lakefs.io/data-quality/data-qualitytesting/

Ma, L. (2021, 09 17). *What is Data Management, actually?* Retrieved from DAMA-DMBOK Framework. Azure Data Ninjago & Dqops. : https://dataninjago.com/2021/09/15/what-is-data-management-actually-damadmbok-framework/

Martin, M. (2023, 12 30). *What is a Functional Requirement in Software Engineering?* . Retrieved from Guru99. : https://www.guru99.com/functional-requirementspecification-example.html

Rabobank. (n.d.). *Data Management Consultant*. Retrieved from Rabobank: https://rabobank.jobs/en/job/data-management-consultant/JR_00082142/

Rachidi, L. H. (2024). *Data quality Management with DataBricks* . Retrieved from Databricks: https://www.databricks.com/discover/pages/data-qualitymanagement

Rad, R. (2019, 6 3). *Create a Profiling Report in Power BI: Give the End User Information about the Data.* . Retrieved from RADACAD: https://radacad.com/create-aprofiling-report-in-power-bi-give-the-end-user-information-about-the-data

Richman, J. (2023, 06 21). *What is data quality?* Retrieved from Dimensions, standards, & examples. Estuary. : https://estuary.dev/data-quality/

Rome, P. (2024, 03 24). *Perforce*. Retrieved from What are Non Functional Requirements — With Examples. Perforce Software. : https://www.perforce.com/blog/alm/what-are-non-functional-requirementsexamples

Sheldon, R. &. (2024, 03 04). *Data quality*. Retrieved from Data managment :

https://www.techtarget.com/searchdatamanagement/definition/data-quality

Simplilearn. (2023, 11 17). *What is Requirement Analysis.* . Retrieved from Simplilearn: https://www.simplilearn.com/what-is-requirement-analysis-article

8. Appendices

8.1. Appendices A: Align dumpstore

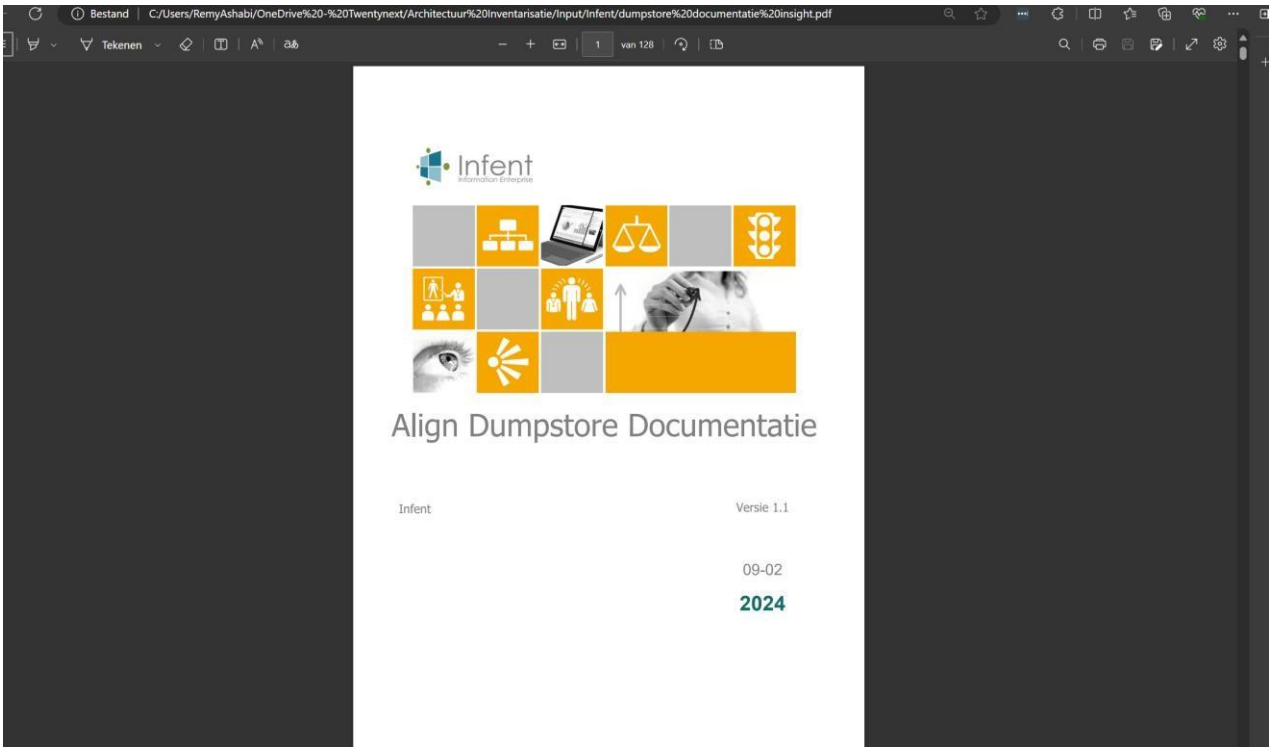


Figure 28 Document 1 provided by AxionContinu

8.2. Appendices B: Kubus

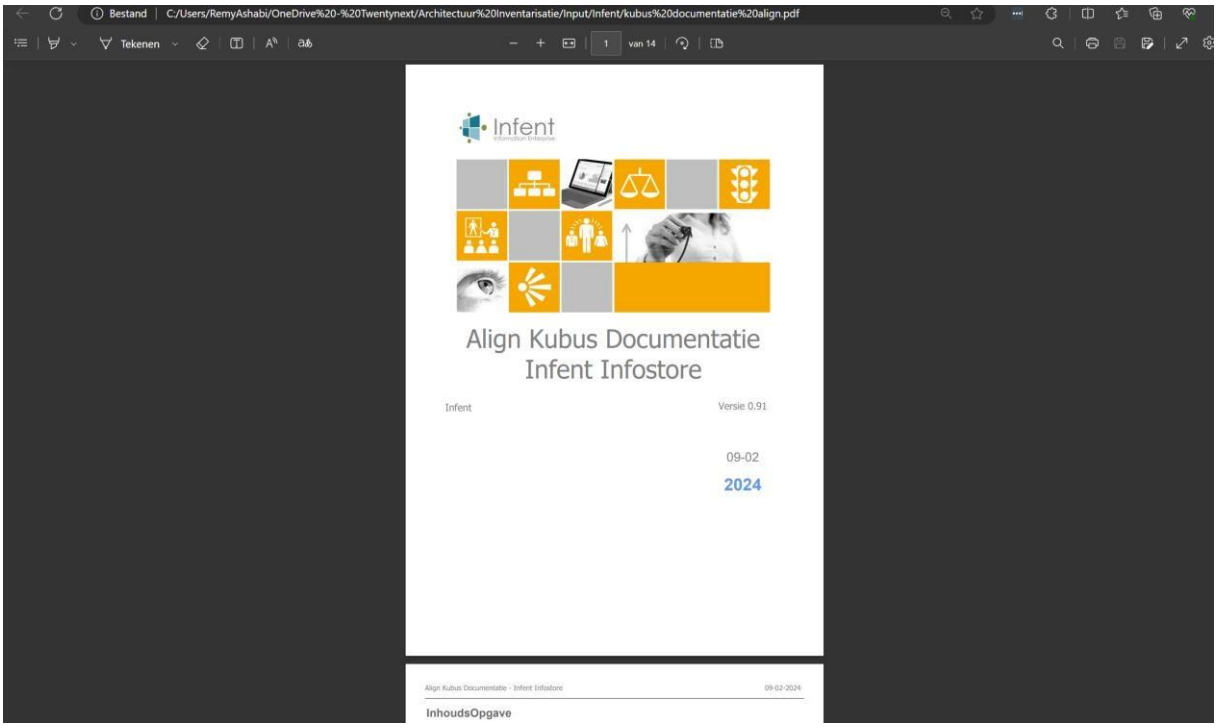


Figure 29 Document 2 provided by axionconitnu

8.3. Appendices C: Align infostore

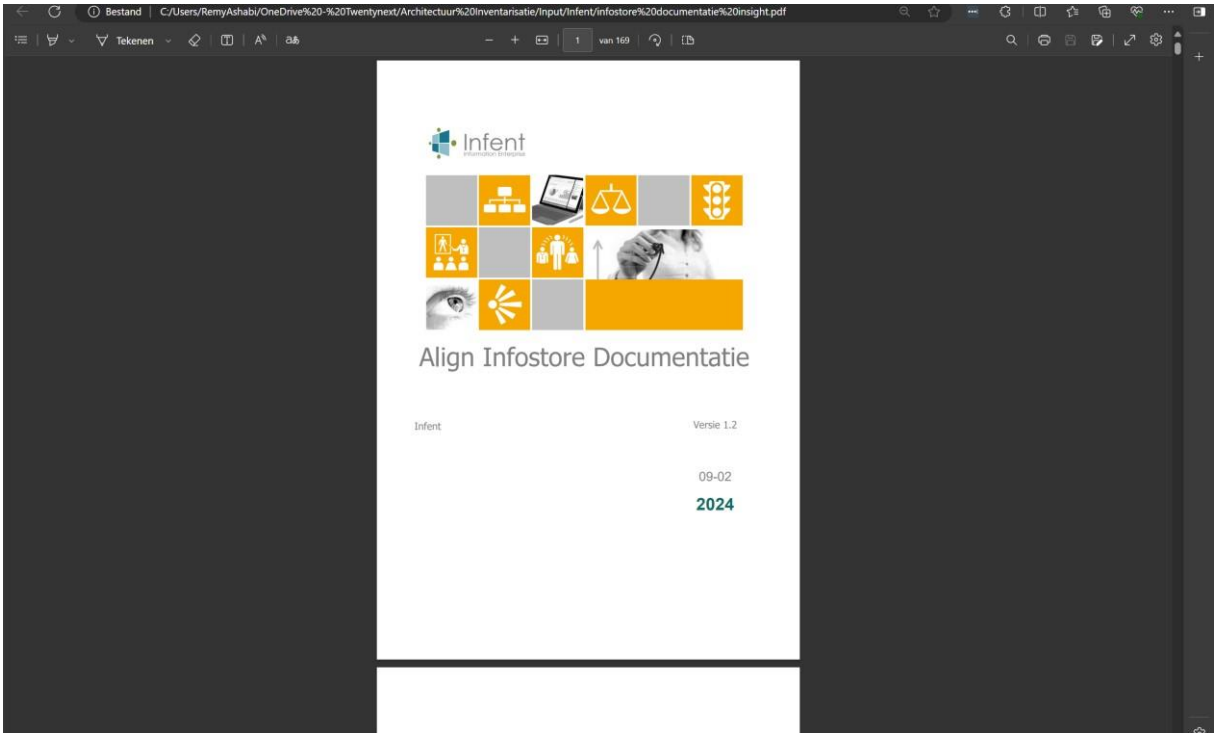


Figure 30 Document 3 provided by AxionContinu

8.4. Appendices D: Interview NI Kang

Audiobestand ni data scientist.mp3

Transcriptie Remy:

Yeah. So the idea of the whole. Project is that I'm doing the data quality part of the project right. I ensure the data quality so I could for example create matrix metrics that cheques the data quality. For example within the consistency or sorry within the accuracy of the data from what I was reading is that we could. Example include constraints to the data, like for example a BSN number has 8-9 characters right? So we could add a constraint for example in pulling the data. Or it's like if it succeeds the nine or something flag it.

Ni Kang:

Yeah. Hmm.

Remy:

If it's less than 9 for this BSN column flag it you know, but don't add it, for example to the main database, etcetera. So this is the type of like. This is the base idea of my project, just data quality and from what I know about you is that you're a data scientist and you are more focused on machine learning. Etcetera. Right to demo, all right.

Ni Kang:

Not focusing on measuring everything about statistics.

Spreker

Could you tell me what you do?

Ni Kang:

Oh, it's really difficult. Guess my I said it intuitives it's computer science, but it's a more empirical research. It's kind of cross section between psychology and user interface.

Remy:

OK. OK.

Ni Kang:

Yeah, it's more like collecting data and then doing some statistical analysis.

Remy:

Yeah, like just your normal distribution, etc.

Ni Kang:

Or hypothesis test. Yes. Yeah, I have done quite a lot, yeah.

Remy:

Hypothesis. That's. Yeah. Yeah. Yeah, that's cool. Yeah. I've done the same in AI because I have an A specialisation in artificial. And OK, we kind of did the same thing. One of my projects was like, you literally just analyse it, etcetera. This is interesting. Yeah. Yeah.

Ni Kang:

Yeah. And then later when I work here, I begin to work on machine learning projects. Deep learning projects is on images.

Spreker

All right.

Remy:

Did you find any similarities in between? Both. Yeah. Yeah, yeah.

Ni Kang:

Of course, if you can't, if you don't really understand the basic statistical analysis, then you don't have good results in machine learning or deep learning. Because sometimes if you forgot like.

Spreker

Vision, yeah.

Ni Kang:

Well, it's basically still statistically analysis, but in a more complex way. But you really forgot those assumptions or the distribution is not good enough, then you can't get good results from the different emotional. Those kind of thing. Yeah. And then I've done quite a lot of.

Remy:

MHM. Yeah.

Ni Kang:

Differently, but I also have done some also time series analysis.

Remy:

Ohh yeah.

Ni Kang:

Yeah. And now also demand forecasting, but this is more complicated. It's the situation because the the supply is never reached the demand level because it's a fresh product.

It's kind of like with everything.

Remy:

Yeah. Yeah, yeah. Can a same thing that happened with Sony two years ago or. Yeah. Yeah. And where when they released the PlayStation 5, for example. And it was like the demand was higher than the production and then they stopped producing it for like, a year where you couldn't get PS no matter what. And if you want to buy one, it's double the.

Ni Kang:

So. This was the kind of thing.

Remy:

Nice. So you see.

Ni Kang:

Yeah, but it is different because this is bread. If they really supply too much, they need to throw it away.

Remy:

OK. Yeah. Yeah, fair. Enough. Ohh, that's Brad. Yeah, yeah.

Ni Kang:

Yeah, that that you never see the real demand time series, yeah.

Remy:

I see. Yeah. You never. Yeah. Yeah, yeah. Yeah. All right, all right. But have you ever worked with the data? Quality within your major.

Ni Kang:

Of course, every project, but usually I don't, I don't do this because I don't. Have such kind of. Big structured data sets, but what I want to ask you like who want to do that or what is your purpose?

Spreker

Hmm.

Ni Kang:

To do that is because data quality can't be. Just make sure this table is good and it's good enough because it's also related to the future statistical analysis or what model you are going to use with the city of data. So you can't like if you don't know any model. Spreker Yeah.

Ni Kang:

That's the analysis you're going to do then you can't really give a really good report on your data quality.

Remy:

Yeah, yeah.

Ni Kang:

You know, for example, if it's really want applying to. Maybe just a simple statistical very fundamental analysis. Later they may have like.

Spreker MHM.

Ni Kang:

Very strict. Like normal distribution as assumption or something like that, but they may require less data like for example. Maybe in this case 2030 cases are enough, but if you really want to put a lot of different factors then maybe hundreds, thousands or but you don't know.

Remy:

Yeah. Or so yeah.

Ni Kang:

Also, for deep learning that 20 no, it's not enough. Yeah, it really depends on different cases or what model or static analysis you're going to do later.

Remy:

You know not. Yeah. All right.

Ni Kang:

But. Yeah. And another thing is like data quality can't can't really as a stand alone thing there. You need to also understand the data, the columns, what does that mean?

Remy:

That's a good idea to think of. MHM.

Ni Kang:

Yeah. And what constrictions you need to put there?

Remy:

OK. So yeah, yeah. So you first have to analyse the data. Understand. Yeah. What kind of constraints fits with what? Yeah, that's true. Because, yeah, you can't just create a normal distribution of, I don't know, string of something, you know? No, I get it. That's a good idea.

UM, but it's always different. You've never worked within data quality within health department have.

Spreker OK.

Ni Kang:

You well for health department. It's health department. What do you mean, health department?

Remy:

Health department is because OK, data quality is could be extending to finance department. You know where they care about the quality of the data within the. Finance. You know, like amount of hours, amount of money, income outcome, homes that, etcetera, right. But they only care about these stuffs, right? Uh. Within, for example, health department, I assume that they care about different stuff.

Ni Kang:

Right, with different stuff. Yeah, I actually, I have been working with different healthcare projects for example like, but this is more image pattern recognition like cancer, breast cancer recognition versus image data.

Remy:

Completely different stuff. Yes. OK. Yeah. Yeah, it's own image data, yeah.

Ni Kang:

And also I have foetal foetal health monitoring during labour process.

Remy:

OK.

Ni Kang:

Yeah, that it looks structured, but it's kind of, for example, this mom is the the labour process lasts around 20 hours and the other 172 hours. And this is only two minutes. It's like and.

Spreker Sure.

Ni Kang:

It's really difficult to say like which kind of data. But if you have a structure, but at least I need to know from. Example. What other for example possible inputs and what are the possible targets to predict or just?

Remy:

So yeah, so for example, the whole idea of the whole thingy is basically these are all inputs are done by nurses, you know at the hospital. So sometimes they could miss, you know, misspelling or sometimes they could skip the whole column that they don't fill or something. Or sometimes they would answer a Boolean with.

Ni Kang:

Yeah. Yeah.

Remy:

True or false or yeah of name, you know. Pictures. So the whole idea that where it comes that I have to make a whole protocol ensure like for example when someone later on pulls the data they have to follow the protocol of Oh no. These columns I have to make sure that they will change the format into bullion. These columns we have to actually run a check to check if it's true or false or if it's at.

Spreker OK.

Remy:

You know, and we could like switch stuff, you know? So making the whole protocol of the data quality and how people could. Where people should look basically you know, because no, yes.

Ni Kang:

So for you, the data quality is more more standalone or more general.

Remy:

More general. Yeah. So that's basically the nurses could have later on when they open their dashboard, they have the full data. You know, they don't have the wrong data. Oh, your name is Jacob, but the BSN is something else. You could be the wrong Jacob, you know? Yeah.

Spreker

More general.

Ni Kang:

Hmm.

Remy:

So I have to ensure like for example here also like you see the space in between the ID. Oh yeah, yeah. So it's like oops. It starts with space instead of, you know, like stuff like these.

Ni Kang:

Yeah, but it isn't really because I I found this really general, so it's.

Remy:

Yeah.

Ni Kang:

You can just. Look up something like it's how it should be defined. And how you can define it is missing or is the wrong format.

Remy:

Yeah, yeah.

Ni Kang:

For every column it's down, but it's a, but do you think it's really a difficult any difficulty?

Remy:

Well, and there is not much difficulty except that you don't really much find much information about the data quality within the health department. You could find it easily for, for example, the finance where you should focus on what kind of tests you could do, what kind of etcetera. But within the health department, it feels like, you know, this is very sensitive data like you could mess up or remove one, something you could, you could have the wrong whole results. You know what I mean?

Ni Kang:

Our for example like. Do do they also include something like a test results from blood tests or something like that? Like they should be also in the range or?

Remy:

I think so, yeah, it includes everything. Yeah, I think so. I mean, I don't have access to the whole data yet, but from what I know is all of the data is medical data, so blood tests, diabetes, cancer, name it, everything right, this is what I got. So what I understood because I don't have access to the data yet. So I assume that there is different ways to deal with the data quality within this area. Like there's a different approach to it. How do you even approach it? Do you have to, for example, go and ask them themselves to is this data? Correct or yeah.

Ni Kang:

Well, at least you need to understand. You need to understand well whatever you need to check with some experts or someone really knows that. But you need to know every column. What doesn't mean? It's really necessary. Is it really like a space? Is is good or not? Or you really don't need to. You need to exclude those things. Or is it just the wrong range?

Remy:

OK. Yeah.

Ni Kang:

Like for Peach values, for example, from a blood test, I think like 10 or three. I don't think it's a blood test or something, right?

Remy:

Yeah, yeah. OK. Yeah. So yeah, based on the research.

Ni Kang:

You really need to understand the value the.

Remy:

Of the columns.

Spreker Yeah.

Remy:

And that's a good idea. All right? And. Good. Good God, I had some questions prepared, but losing everything. How do you envision validating now? That's cool.

Spreker Yes.

Remy:

So yeah, yeah, so. Hmm. Yeah, because you mentioned it. Yeah, that's a bit different. Umm. There it is. Eddie, do you use any specific methodologies or frameworks when you check with data quality? Or do you just go based on Hunch?

Ni Kang:

Well, it's just kind of framework. That's what you have sent you, the Christian.

Remy:

Yeah.

Ni Kang:

It's like like for me as a data scientist. I'm more doing this data understanding and you here and the evaluation other things that usually I don't care. But the understanding is really important for me. Yeah. But The thing is mostly I work on the very.

Remy:

Yeah. I see, yeah.

Ni Kang:

Specific context and specific research question, so usually the data quality check is always related to the model I use.

Remy:

Wanting. I see.

Ni Kang:

And first, you really need to understand the data, but this is also general because. For those kind of missing data outliers or those things, outlines also relates to the context itself, right? Missing data and wrong format. That's the really basic thing we need to check like distribution.

Remy:

Yeah, yeah, yeah. Yeah. Do you have? Yeah. Do you have any any specific words? Do you check these, like, how do you, how do you check these? Like, just just counting of null. Or future like also distribution. But what else do you do to check the data? Understanding you know what I mean.

Ni Kang:

It's really depends on different data because I this is my working content is this time is only text and next time is only images and next time is data set data set.

Spreker

OK. Yeah.

Remy:

Edges. Yeah, fair enough.

Ni Kang:

Looks, which looks like a structured data, but actually it's not structured but. Every time is different.

Remy:

Understandable. Alright, but if from what you've seen, for example from the data that I just showed you, right, what what would you advise like? What kind of cheques did I histograms?

Ni Kang:

Histogram such like a histogram, like if they have a definite number of values they have used. If it's a categorical, how many categories and every category how many instances the cases.

Remy:

OK. Yeah. This yeah. Alright. Yeah. In the histogram school.

Ni Kang:

Yeah. And range and also understand the meaning of the column is really a good range in that for that data. Yeah. And also I have working, I I've been working with the vital health monitoring during the labour process.

Remy:

That's.

Ni Kang:

And the data from the hospital is also not good. Sometimes it's just a. Like they they married, they give a score for the the baby after birth, like one minute and 5 minutes and sometimes like 1 minutes 10.

Remy:

MHM.

Ni Kang:

And five minutes later, one it's like, is this really you need? Yeah. You don't know to ask like sometimes because that project was related to the monitoring or the management of the models construction during the neighbouring.

Remy:

It's a puzzle. I see.

Ni Kang:

The labour process so. Like the score after 5 minutes is 1 but 1 -. 10 is not related to the labour process, and actually that case was related to the baby itself. The baby has its own problem, so it shouldn't include that case. So.

Remy:

Oh. Yeah. Casualties and causality basically.

Ni Kang:

Yeah. And also sometimes it really gives the wrong number.

Remy: M.

Ni Kang:

Like a healthy 1, which they gave like a 55 is the uh. It's a really dangerous one, but I don't see any abnormalities in the in the in the monitoring process. Oh, I did the wrong. I did wrong. It's OK. Yeah, you can. You can't. Yeah. Also for medical thing, you also need to say.

Remy:

I see.

Ni Kang:

It's really difficult because if you want to understand the thing and a lot of things are related, they have to do. Some have done some surgery, the medications and the number and all related real a lot of different.

Remy:

M. And a little, yeah.

Ni Kang:

The knowledge to know, but if only focus on the general thing, the format is OK, but the number itself is. Sometimes you can't see directly. It's really a lot of outlier now.

Remy:

That's true. Yeah. Yeah, that's very true. Yeah. That's also was one of my questions is like, like when I was going through the data, how would I know if this number is an actual outlier? You know what, I. Mean yeah, because I have no idea. Yeah, no.

Ni Kang:

Yeah. First off, you can't. You need to consult some experts on this.

Spreker

Yeah. So yeah.

Remy:

Basically, and that's that's basically what we were doing here. So the green we ask basically we send it back and we ask them is this actually true or not? Yeah. All right. I have one one last question for you. Throughout your data preparation methods and.

Ni Kang:

Yeah.

Remy:

Data understanding parts. Have you ever came across any challenges or? Yeah. Have you ever came across any potential challenges during the data preparations and data understanding that made you go, huh? And how did you? Approach it.

Ni Kang:

Did the understanding for me the most challenges? Well, it depends really understand the data. For example, I have done the the breast cancer detection that part and in the data set they said I have annotated the cancer area, the tissue from the tissue.

Remy:

Yeah.

Spreker M.

Ni Kang:

And the other colleague, actually the other, yes. And OK, then I crop out all the tumour tissues as a separate set as the unhealthy one. And they have the all the sets without.

Remy:

Yeah, yeah. Yeah.

Ni Kang:

But actually what they annotated is they annotate the area is really cancer tumours. On the same images on the same image, they still have other tumour cells, but they didn't notice.

Remy:

Ohh. I see.

Ni Kang:

And also when later I saw when it annotated the bonding at the edge of the tumour area and sometimes it just goes through.

Remy:

Yeah.

Ni Kang:

You you really well, they looked exactly the same, but you say this part is tumour, but that part is no, it just goes through, OK.

Remy:

All right.

Spreker

It's.

Ni Kang:

Then I really need to like study a little bit the pathology. They said normal cells, really normal cells and this is look like a cancer or tumour, but actually no. But these are really the real tumour cells you need to.

Remy:

OK. Yeah, yeah, I see.

Ni Kang:

Understand those kind. Of things.

Remy:

So it made you go back into the basics of the biology to understand, yeah.

Ni Kang:

Yeah, why the model doesn't work? And so because the data set is not good or I don't understand the data. Another challenge is sometimes, like the data wallets, the assumptions of the potential. Best model you're going to use. Then go back and forth. I want to use that model, but this assumption is violated, so I need to change another one, but I can't find another model, but they are really related and those are really difficult thing.

Remy:

OK. Spreker

Yes.

Remy:

I see. Yeah, I can imagine especially with images and stuff, it's a bit annoying. All right, all right. But yeah. Yeah. Well, one good question from what you've seen like imagine a categorical data, right? How could you define to you right, what is a good data set, what is? Yeah, OK now how to reference this? From your own perspective, or from your own expertise, what is a good data from a categorical data? How would you define a very good data that data quality in this data is good? How could you give it? Yeah.

Ni Kang:

It really depends on the context, like how we're going to use that category data.

Remy:

OK.

Ni Kang:

Well, at least from the very basic quality issue then it should be. For example Catalina Sickle House and Catalina Hospital. They are the same thing. They mean or not.

Spreker

Mm-hmm.

Remy:

Yeah, I see.

Ni Kang:

Yeah, they sometimes they use different name, a little bit different, yeah.

Remy:

Consistency is basically, yeah, yeah. All right. Is there any other stuff that you think that actually makes the data has quality or it's?

Spreker Yeah.

Remy:

Like. Yeah. You get the question.

Ni Kang:

Yeah, quality. If they're not missing and the value is correct, at least in some extent should be correct and.

Remy:

OK. Alright, but if you don't know if if they if it is correct, you know. What I mean? Can you? Yeah, that's.

Ni Kang:

You can.

Spreker

You should just go and ask.

Ni Kang:

All right to ask like, well, if you just one or two outlier, you can maybe it's not really you can just discard. But sometimes for example. And to detect the bad cases from the the vital motoring. But the bad cases is only like 10 cases out of 2000. I can't really discard those. Then I can't do anything.

Remy:

I see. Yeah. I know. Yeah, yeah, yeah, fair enough.

Ni Kang:

It really depends.

Remy:

All. No, no, no. I think we have covered everything. Thank you. Yeah, it's it's really general for now because it's it just started, right. It's really the basics because they don't have anything for data quality.

Ni Kang:

Gives a really general.

Remy:

So it's really just to start trying to get different perspective from data scientist or because you said you've worked with machine learnings now at this moment. So I want to get your own perspective of data quality from a different, completely different side. I also will get Michael's ideas.

Ni Kang:

Yeah, I think this is more this quality is more medical.

Remy:

Yeah. I know I also like don't have the don't have access to the data yet, so I really can't do much into it, you know so.

Ni Kang:

So what does this system is going through? Is it just the?

8.5. Appendices E: Interview Linda

Audiobestand

Linda interview.mp3

Transcriptie Remy:

Hi, good morning, Linda. Good morning. Yeah, I was wondering about your expertise. So could you tell me a bit about? Yourself.

Linda:

Yeah, I'm Linda, im 29. I work as a senior data engineer here at Twenty next. My previous experience was as a data engineer mediator at Rabobank. And besides that, I. I did other roles in tech. Like Technical Support analysts, where it was more on the infra side also worked as a data analyst. That's why I also understand more the business and the requirements of the business and I even worked in orbit like I was in it all the time. So I was I was auditing.

Remy:

Jesus. Auditing in what though?

Linda:

Also like data quality. Like like identical and access management controls with access to systems. So I look also at the source, you know? Yeah, because as a data engineer you extract data from the source and when data from the source is trash.

Remy:

Of course.

Linda:

We need to do a lot of work to clean that data and then to add some new columns or etcetera. What the business needs. But it's all started by the source and the the source. I mean like for example when employees fill in their cell phone numbers, if they're not, if there are no input controls in the field or have to.

Remy:

Yeah.

Remy:

Like requirement required.

Linda:

You have to fill in. Yeah, the input controls like cell phone number needs to be at least nine numbers. If it's not there, they can fill in anything. And if it's not like must be a string, then they can fill in anything, and that's already the the the.

Remy:

Yeah, yeah.

Linda:

You know, so the first thing you do when you extract data from source, you do data profiling to see like, OK, what data is there and what is the yeah quality and percentage. And you can look by yourself too. Are there a lot of new values? How is it there you know? And based on that you will actually make for yourself. Try to see like OK when there are no values, for example into this primary key. It's actually mostly of the time the primary key cannot be null, no. So when a primary key is null, do you change it into? No, instead of like, no, you know the end. You then the no like 0 indeed like that you find out. OK, what do we do? And you communicate with the business. Is it OK? Like if there's no integer 0 then we let them zero or do we just remove them from the list. So that's for example.

Remy:

Yeah.

Linda:

And it's also doing the data engineering. The like the there's you must check the data quality and one of the thing of data quality is the data accurate. So it's you can build a check like if the the data the source date is. Smaller than yesterday, for example. Then you know the data is not accurate, for example then.

Remy:

OK. So if the date is smaller than yesterday. So yeah, yeah.

Linda:

If the source date is the source date in the file is.

Remy:

Yeah.

Linda:

Before. Today, yeah. Then the data is not up to date today, yeah.

Remy:

Oh, today 20 today. Yeah. Yeah, because might be yesterdays or so.

Linda:

And there might be many reasons might be that the person who needs to provide the data from the source system did not provide us the data on time, or it might be other connection issues that when you were performing the the trigger, you know trigger that you say like OK, this time the pipeline has to run, maybe that trigger was just.

Remy:

Yeah.

Linda:

2 minutes too early, before the data was ready. You know, it might seem any reason, so data actuality is a thing you look at.

Remy:

Yeah, that's fine.

Linda:

You'll also look at. Yeah. How many? What do you do with null values if there are no values, what do you do with?

Remy:

That's a very good question.

Linda:

Yeah. No fellows, that thing and what do you do when the data in the? In the fields are not. That you expect? Like what?

Remy:

OK.

Linda:

Do you do when in? You have colon bone plants where someone lives. When when you see their ABC D.

Remy:

OK.

What ,it's not supposed to do this.

Linda:

Those things, yeah.

Linda:

And the data profiling you can look in detail like OK, these are the columns. These are the attributes and these are the expected values and if you see like very strange values, what do we do with them? Yeah, mostly of the time you change them or you.

Remy:

But how do you know if it's actually accurate when you changed it like, but you have to 1st ask the business, yeah.

Linda:

Yeah, it fits as the vision like we saw in 10 cases. That these records are not like you expect. In case if we look at the entire table, we see these specific records deviate. OK? Yeah. Outline. So what do we do with them? What do you want us to do? Do you want us to take in the values of these missing records there. Efforts.

Remy:

Yeah. Yeah, outliners.

Linda:

From all like the average price. For example, if the price is missing.

Remy:

MHM.

Linda:

Or do you want us to eliminate these records?

Remy:

Yeah.

Linda:

That and when you know what the business wants you do, you do, you do what the business wants because at the end, if you make your own decisions without talking to business, the business will be angry because, yeah, you did not do what you wanted. So now we don't get the reports. Now the reports are not right sometimes after release.

Remy:

You can. That's always like, what are you doing? Yeah.

Linda:

Change the structure of tables and stuff and then the business. When they look at the power BI reports they were saying like. Well, this, this, this, this value is not right. It's not. So then we look back and then like, Oh yeah, we must have changed something and the stored procedure, the stored procedure we used to insert new data on an incremental basis.

Remy:

Yes. Let's see. So something went wrong getting down? Yeah.

Linda:

Yeah, like long in the stored procedure and that's why the business business logic was a little bit different. So we have to adjust the business logic in there and then everything in the power BI reports was right. So it starts actually with the source.

Remy:

Yeah. All right.

Linda:

And the love of garbage in the source. Then you must consider like, are you going to use this source or are we first going to set up the input controls so the people who produce the data produce correct data and it also has to do with the training of people and. Spreker OK.

Linda:

And you have a new system. You need to train people like OK in this field you you use this, but that's the people you have people you have processes and you also have systems. So in the systems you can already put controls. That's actually one of the best thing actually because then you yeah you enforce that it's not able to.

Remy:

For both required controls.

Linda:

Put in other values.

Remy:

Yeah. Yeah, yeah, yeah, I see because.

Linda:

Mandatory.

Remy:

Basically, yeah, we have jumped through the whole data life cycle. It's like, yeah. We could split it up and and in case of action continue right, like during the data collection phase or the data input phase. We don't have much influence on it, right?

Linda:

Data input from the source from the source itself.

Spreker 3 Hey.

Remy:

Like we don't have much influence on this. When the nurse enters the data, we do not have any influence on the application they use for that, right?

Spreker 3 Yeah.

Linda:

No, we don't. Yeah, we just get the data. Yeah. And yeah, the, the the supplier of the data is actually responsible of what they deliver to us. And we are responsible of making that data usable for the business in reporting. So yeah, we cannot do much.

Remy:

OK. Yeah, yeah, yeah.

Linda:

About the source, but we can discuss. The OR findings on the source to the business.

Spreker

We could like give.

Remy:

Them a recommendation and apply so that's yeah.

Linda:

Yeah, like, OK, I see this. This column is all.

Remy:

That that's a very good idea, yeah.

Linda:

Actually data. So can you make input controls so in the future it won't be so you can erase them the the historical data later and then use only the future data which is more correct, yeah.

Remy:

Yeah, and OK, that was during the data collection phase.

Spreker 3 Man.

Remy:

Do. Yeah. During the entry level, right? How do you ensure the consistency and standardisation in the data entry process? Yeah, we mentioned that where we did add required fields and the trainings, OK. And during the storage level.

Spreker 3 Then.

Remy:

Because I'm guessing, yeah, they also storage but.

Linda:

Storing the data that is extracted from the source, yeah.

Remy:

It's. So yeah, from the source, right. So do you previously from your previous experiences, do you all have you used any strategies that?

Spreker 3 Yeah.

Remy:

Any strategies that you have done to ensure the data? Integrity and security.

Linda:

The the data integrity when we extract data from sources we also apply hashing strategy and this hashing. You have the data source and then you add a column with hash value and the hash value is created by a combination of the entire.

Remy:

OK. OK.

Linda:

The row and then you know the uniqueness uniqueness of each record. So not only primary key but also the hashing. Hashing is also. Then you also in the hashing is actually a concatenation of all the columns per row. So you have a column hash value, and that column exists of concat of the date where the data is sourced. The date time in seconds and stuff and all the other things. So then you can always look back like OK this each record is unique, really unique and you see that this.

Remy:

OK.

Linda:

This data was from this date and these seconds, so that's how you maintain the integrity and also what will be more also the source data on that day. You can always compare it with the data that you already have stored.

Remy:

OK.

Linda:

By writing a query or script like this is the source. This is the and and then you can do the except you can test it with except this this source and this is the data we store it and if you see no records coming out then you know yeah, this is exactly the same.

Remy:

Same alright, but then this is another good question like because OK we pulled the data from the source VM API or so and we tried to pull it one-on-one right where it's like whatever it was we tried to make it whatever it is.

Spreker 3

Yeah. Yeah. Yeah. Right.

Remy:

But sometimes this doesn't go as wished, right? Or something goes wrong in between the whole process.

Linda:

Yeah, in from source to destination, yeah.

Remy:

Yes, from source to destination right? Like although we tried to make it one-on-one right, but how do you ensure or how do you yeah. How do you ensure that that transition is 1 to one you know or is that a metric that you installed or have you done before?

Linda:

The transition is 1 to 1 low in the Azure data factory for example, you have a copy test. And there you really specify like this is the source and you can also preview the source and you also define the sink. The sink is the where it's sync and there also. So it's oneon-one net like in Azure Data Factory and then.

Remy:

OK. Yeah. Yeah, yeah, yeah, that's true.

Linda:

Seeing source sync and then you also have a preview of the of the sync. You can also see like OK this is how it looks like when the data is synced and in between you can also build that check.

Remy:

Source. Yeah. OK.

Linda:

So like this is the source and if the source is empty then of course there's no sync.

Remy:

Yeah. Yeah, yeah, yeah.

Linda:

And then we can get the e-mail like, OK, there's no new data delivered from the source. So this is, yeah, there's nothing is done. So there's a one-on-one mapping and when you make a data flow test, you can also map. That's also the good thing that you have.

Remy:

So that nothing is done yet, yeah.

Linda:

Comma and then you map like each column from the source with each column from the sync. Yeah. And then you know that the mapping is good. So yeah, yeah.

Remy:

Film. Yes. Yeah. Such a lake house. When you try to. Yeah. Yeah, but OK, I I yeah, I had another good question.

Remy:

I had a very good question, but it just left my mind. Although we map it right and we still sync it and we, yeah, we do the same via Azure Data factory, but I've seen in one of the records that Jasper colleague showed me is that even after the data has synced, there's still some errors in the column itself, like extra variables for example.

Remy:

Right.

Linda:

Errors in the columns itself. Oh yeah. Yeah. They still, yeah, it depends actually, because it depends how the process is between source and sync. And then in this case we only like get data from the source and then sync it and then in data breaks.

Remy:

By for example, yeah. OK. Just.

Linda:

The offer the, the, the, the the cleansing. Like, OK, let's do things with the arrow. Yeah. Sylvie does that so.

Remy:

Oh yeah, in silver.

Linda:

But you can also do it in Azure Data factory that those cleansing things so that only the good version is synced. Yeah, but we they decided to do like only the the data bridge that they do there the the the ECL process actually more, yeah.

Remy:

Well. Hold on. ETL. Process. Yeah. Yeah and.

Linda:

After we extract the absence in Azure data factory and then the transform in data.

Remy:

Yeah, I was reading about this too, because one of the best practises that was mentioned, I don't remember where which source, but I think it was Microsoft.

Remy:

You mentioned, yeah, if you followed the data break Medallion framework, then it's better to just drop all raw data in bronze, no changes, and then you do the detail in silver. I guess that sort of approach, but then the silver part would be considered the processing level within the data.

Remy:

This life cycle. OK. And. Yeah, we talked about how do we ensure the clinical data, the accuracy and the reliability of the data where we do the analysis and then check with the business depending on the business.

Linda:

You can build like the data actually. Actually type check so there is a column and then you have you add an extra column.

Remy:

Yeah.

Linda:

Actual and then it's a bit like a bit.

Linda:

All right. So the one if it's one, it's two and zero, it's false. So every day when new data is loaded, it stays at 1:00.

Remy:

Is. Yeah. OK.

Linda:

But if there's no new data loaded, the data is from yesterday, it goes, it goes zero. And then there's, yeah. And when it's zero there.

Remy:

So everything.

Linda:

You can automate it to send the e-mail like with the list of the options which are not actual and that's how you can maintain the actuality of the yeah.

Remy:

That's a good one. Yeah. Yeah. These are good majors.

Linda:

These are very good metrics that I have also at the riverbank. But now first my first assignment here is to build a basis platform. So first like. From A-Z and then later the cheques for data quality because that's very important actually data quality, otherwise you cannot make informed decisions.

Remy:

Shouldn't shouldn't the data quality go like by leg with the building of the whole process? Yeah.

Linda:

Yeah, that's how I think about it too. And next week I will discuss it also with Mike. And Dali and the Oscar. But that's that's very important to at least check the data actually like is it actual is there new data if if not we can talk to the business. Hey, there's no new data. What's going on? Can you deliver the data later so we can trigger the pipeline. So the business has data so yeah.

Remy:

Yeah. Thank you. Yeah.

Remy:

OK. And yeah, missing and incomplete patients data. Yeah, as we mentioned earlier, we checked with the business and mostly most important stuff is that the primary key should never be null.

Spreker 3

Yeah, yeah, that would be enough.

Remy:

Yeah. Are there any automated tools or algorithms you use to detect the correctness or errors of the data?

Linda:

The errors here, for example in the pipeline you can also build. The script to check the arrows in the data arrows means like you define the arrows first. What do you see as an arrow? If you see that an entire column is empty, for example, then that might be seen as. You can define all the errors that you consider an error and also with this is you can talk about OK.

Remy:

OK.

Remy:

Thank you very much.

Linda:

What do you see as a fault? What? How do you define false and then you can make a task actually with the script in it and that once before the data will be ingested into the delta lake the block storage.

Remy:

Yeah, yeah. Yeah.

Linda:

There are many steps you can implement eval check up to check if you if you see that the data type of the source is not the same as the data type that we used to ingest those things and then we need to talk to the business. Hey, you see that the data type changed but you.

Remy:

The teacher.

Linda:

Did not communicate. Yes.

Remy:

Well, like if the data type changed is that comes from us or from the business part because.

Linda:

It can be from both child, but we don't change the data type without the consultation of the visions, but most mostly it can happen that the business change things or they change like the the source, the structure of the source so they remove some columns or add some columns and we didn't know.

Remy:

OK, now that's consulting. That's true.

Linda:

From that and then things can also go wrong in our pipeline because we copy 1:00 and 1:00.

Linda:

But we also based on what we copy, we create dimensional table 10 tables, fact tables and if things in the source change without them knowing, we cannot change the thing from the entire process and then pipeline will fill in.

Remy:

You know. I remember now I have a similar. Problem in my previous experience where yeah, business changed the whole source and then everything fails and it just depends that the API just changed their names and stuff. So I was like.

Linda:

Oh yeah, yeah. Also the API exactly.

Remy:

Yeah, yeah, yeah, it's no wonder this is.

Linda:

No wonder there's no data then.

Remy:

Fading for the whole day. All right. Are there any measures that or measure? Yeah, matrixes that you have used for the accuracy and reliability of the data within the processing level?

Linda:

Forces and accuracy. Yeah, we we checked the actuality. Yeah. So you can also build a based on the metadata. You can build a table.

Remy:

Yeah, which?

Linda:

Where you check actually the actuality. So where per source you see OK, this is the source date and this is the date of today. If the dates there's actually a flow. If these date is not today then not actual send an e-mail or put it in the dashboard mostly emails.

Remy:

Yeah. Yeah.

Linda:

Effective because they'll negotiate, we have to. Do something. Yeah, that and the completeness of data. It might also be that duplicates, yeah. So you also check like the entire table having, yeah.

Remy:

You need less time games, etcetera. Yeah, yeah.

Linda:

Having counts more than one and then if that is so, yeah, then we need to in the source. We need to remove to. The application.

Remy:

Yeah, fair. But how do we know that? This is an. Actual duplicate. You know what I mean? Like. Yeah, we could check and identically, but you mean? I mean, like, how do I know if this wasn't intended on purpose to have him in a separate line?

Spreker 3 Yeah.

Remy:

Or something, or someone messed up.

Linda:

Yeah. Why would you have the same records 10 times?

Remy:

And that's true. You catching.

Linda:

So of course you communicate with business. Like we see a lot of duplicates here and mostly duplicates are not fine because when you use your stored procedure to insert data you will.

Remy:

Yep.

Linda:

Get an error like. The merge could not take place because we cannot update. There were more than twice, so that goes wrong, so duplicates in the source must be. And I don't think there is a reason why you should have the same record 10 times, like the same everything the same.

Remy:

Alright. Yeah, yeah. No, I didn't make sense. Honestly, I 100% agree no.

Linda:

Otherwise we cannot do. The work for you, so fix the source, yeah. And we can also fix it because we take the data and then we make a check to check if they're duplicates. We can remove these duplicates. We can do it, or they can do it themselves, which is actually better. They do it themselves. So we have less times tomorrow, yeah.

Remy:

I think it's like organised. Yeah, OK, yeah. But let's talk to the integral. Let's talk about the integration level, because from what I got till now is that from dialling the others is that we're focusing now on Minecraft is.

Linda:

Yeah, the most important source for now.

Remy:

That's. Yeah, that's yeah. But later on, it's coming up files. It's my etcetera, right. But then how do we?

Remy:

So yeah, how do we integrate the data from multiple sources while maintaining the quality?

Linda:

Well, we we want to use for data ingestion. We want to for each source we build a pipeline.

Remy:

There's also different times different.

Linda:

So for for one source you can have for one source you can get like 10 tables. Yeah, that's OK. The list of so per source a pipeline and and for each pipeline we almost adjust the same strategy, but it depends also on. How much the data is per source like 9 cars? If it's like 120,000,000 records then you need to make a difference. Trying to see how how you're going to load the data. Like do we load it like in parts like in the morning 20 million and then 20? Because if you do it at. One time it's. It might not go well. There might be a timeout and so so per data source we create a pipeline.

Remy:

Which one is?

Linda:

With all the cheques the same uniform like the others, and some more cheques. So per per data source of the pipeline, actually, yeah.

Remy:

So Berlin. That's OK. Very good one.

Remy:

Yeah, I don't think we have any data governance governance principles right now, right. Yeah, yeah.

Linda:

Yeah, yeah.

Remy:

Yeah, I think it's gonna be also part of my project where I'll have to make a whole data governance.

Linda:

Yeah, that would.

Remy:

But I don't know how. Deep. Should I go into it? Because yeah, I could still make it very vague going through all the dimensions of the data quality like, yeah, it's just governance of a person should do this and that and that, yeah.

Linda:

Yeah, yeah, yeah, indeed. When this happens. Like, what do we do to maintain the data quality? And it is about the people. It's about the processes and about the system.

Remy:

Yeah, responsibilities to roles and responsibilities.

Remy:

Yeah. All right, this is very good. I don't. Yeah, their presentation level. I'm not going to talk deep into the data presentation level or the usage because that's not part of my scope, but there.

Remy:

Is part of. My scope of data presentation which is representing data. Quality in the dashboard. All right, so.

Remy:

I think it will be a very nice to have it in a different meeting where we will go through the different data quality like uniqueness standards etc. And then we could define matrix for every one of them that we could use for the dashboard. Yeah, but that will come into later on.

Remy:

Let me check. Yeah, data maintenance level. Yeah. What strategies do? You implement to maintain data quality over time. That's. A very good question.

Linda:

The the quality over overtime, I think it's more for me a daily thing when I'm doing my work and and you can also build a maintenance pipeline and then the maintenance pipeline you may maintain the sources that you take in and then you check for example.

Remy:

Yes.

Linda:

For all the pipeline and where you check for all the sources like. Is the data accurate for example and check is the data how many empty records are there? Yeah, now is the. Are the sources used because sometimes we have sources like fuse or tables that are actually not used for here. So you can think that OK, maybe we should.

Remy:

I don't know. Yeah.

Linda:

Delete them first first, put them in an archive. Shame them because sometimes you delete something and then it goes on and then later delete those sources, because sometimes we have many tables and stuff, but you don't.

Remy:

Yeah.

Linda:

Use it.

Remy:

Yeah, that's true. That's true.

Linda:

And then we can discuss with the business like we. Based on this query that I wrote, I see that this field is not used for two years. Do you still need it? Yeah, if not.

Remy:

Is this documented by anybody?

Linda:

Here it's it's. It's in my head from my previous job.

Remy:

No, it's the same because I in my own experience too. Where we document these types of stuff, right, like so we get a table we document per table all the all the columns next to it will be like there's a tool lifting. Notice this has to be used for long. Should we still take it? Etc.

Linda:

Exactly. Yeah. Yeah. They can send like an e-mail. Like, OK, check.

Remy:

Yes, well.

Linda:

Check the fuses that are not used for a long time. Check the fuse. The broken fuse for example. Check that, that's just something I did it once in two months. Each colleague had to do those maintenance tests, so that's why you have a data lake data warehouse with accurate data and data that.

Linda:

It's actually used.

Remy:

So from what? I'm getting right now is that data quality is everybody's responsibility.

Spreker 3 Hey.

Linda:

Yeah, actually it is. Yeah, yeah, yeah. People should take that in their job, not only making new stuff, but also maintaining the stuff that is made. Yeah. Yeah.

Remy:

When it works. No. All right. I mean, I'm literally almost done with all the questions, right, because. You've already mentioned asking questions, which I love honestly. Spreker 3 Yeah.

Remy:

But I mean. As we're early, we could already jump into the metrics of the data quality per phase perfect by me. So we know that in data quality we have like 7 dimensions, right? OK, let me open some of them.

Spreker 3 Yes.

Remy:

Uh. Yeah, we have 7 dimensions, right? We have the accuracy, completeness, consistency, validity with and integrity, right. So we mentioned accuracy. We could include column of the zeros and ones. Yeah, yeah, if anything.

Spreker 3 And.

Yeah.

Remy:

Before today.

Linda:

Yeah, everything before today, we know that correct send emails. So we can figure out why what happened with the source did they didn't they deliver the source or maybe sometimes also a developer created a batch script and that's why the source is not getting there. So there must be more reason but then we can find out and and fix it.

Remy:

It's not accurate.

Spreker

Yeah. Thanks.

Spreker 3

Hmm.

Remy:

And the completeness. Is that we could find the ratio of empty values with the whole rows of the whole table and then we could always ask the business back.

Spreker 3 Yeah,

yeah.

Linda:

Is it complete and why? It's just zero. Why? That's yeah, that's yeah.

Remy:

Consistency.

Linda:

And completeness is also like the the duplicate. Yeah, yeah. Complete.

Remy:

Isn't that uniqueness?

Linda:

Yeah, uniqueness also. But if you have like a a table of 50,000 records and 20 thousands are.

Remy:

Duplicates.

Linda:

Duplicates there.

Remy:

Yeah.

Linda:

It's yeah.

Spreker

Yeah, I I get.

Linda:

What you or or maybe like the completeness is also like, what if there are philtres down like where? Where the year is between 2000 and 2025, then you actually exclude some data. That's also probably, but that's not in this or maybe it can be in the source that they exclude some data, maybe it can be us saying that we only use the data of two years. So yeah, but.

Remy:

True, that's not yeah. Yeah. Even if we did exclude the data, for example, of two years, we really have to have a good business reason.

Linda:

Yeah, business reason why you only want two years. Do they want to only to see the reports for the date of two year? Yeah.

Remy:

Yeah, well, yes, since this is healthcare department, I think it's very mandatory to have the data from the staff right, because you never know. Yeah, history is very important, I guess. Yeah. Yeah. OK. That's the completeness. Consistency. How do you measure consistency? That's a very good question.

Linda:

Yeah, yeah, from the start.

Spreker 3 Yes,

Sir.

Linda:

Consistent, consistent is that, that consistent every night new data is loaded or.

Remy:

I think so, but that's also timeliness. No, but the timeliness is, how do you time the data from different sources, right?

Spreker 3 Yeah,

yeah.

Remy:

Yeah, consistency would be ohh no consistently if something like how do you. Here like for example Boolean of yes and no. Some people would insert yeah.

Spreker 3 Yeah,

yeah.

Linda:

Yeah, yeah. Of the yeah. There must be consistency. Indeed that we just use Boolean. That's what I prefer the most. Like. Yes. No, just boolean. So.

Spreker 3 Yes.

Remy:

But how do we know if this colon is actually a Boolean column and not just a string column where?

Linda:

That's why in the beginning, when you do data profiling. You see from the source what data types it is. Yeah. And then you think, like I see sometimes. Yes. No. And I see because sometimes the data are so you can also fill in a zero and no and then yes, no issue. Yes NO10 like what?

Remy:

Yeah.

Linda:

What do you want? And then, yeah, and then you talk to business and we can also standardise the entire column that is.

Spreker

What's going on?

Linda:

Bully. That is, like, bully, you know, and that's the 10 that you do that actually and later if you can put it 10 first and then say OK we want yes and no. But it's first two up to the business. What do they want zero or no or true or false.

Remy:

Yeah, yeah. Calls or yes or no.

Spreker 3

Yeah.

Remy:

Yeah, that's them. That's very true.

Linda:

There are no in power BI when when they get data from the data warehouse which is. Bullying. You can add to false via of online, so that's also.

Remy:

Yeah.

Linda:

But it depends what the business want to. See in the report sector.

Remy:

Yeah, yeah, that's better. But is there an automatic metric that could measure the consistency?

Linda:

The consistency of or if if they're, if in the column only bullying or yeah.

Remy:

Table. Like, yeah. Is there something automatically? Automated. That could, yeah. Measure that.

Linda:

Should affect that the values in one column if it's.

Spreker 3 Yeah.

Remy:

So, but we'll have to do it per column per table, right? If that's.

Linda:

Yeah, let's check if this column has is and yeah, the data type is fixed.

Remy:

OK. Yeah.

Linda:

That is fixed, but only what is in the data. But if we work with bullies, it's zero or one only that's a cool thing. But when you work with the far charge, you can literally only feel.

Remy:

David.

Linda:

Everything, all the characters strings, you can feel anything in there? Yeah. They oh, yeah. With AA or. Yeah, yeah. That's why they prefer to.

Remy:

Yeah. Keep it standing.

Linda:

Yeah, if it's and all you need in the source, you need to enforce that you only can choose like in drop drop down. Yeah.

Remy:

Yeah, but that would be very good. Best practise of stuff, yeah.

Linda:

That's awesome. Yeah, if they want. Yeah. And they then want to see in the source. Drop down this. Yeah. Not that people can type it in because the more people type in, the more human errors. Yeah.

Remy:

That's very true. That's very true. Validity. How do we measure validity?

Linda:

And let me think solidity. I have to think so long ago that I eventually died.

Remy:

Hey. Hey. Dimensions. Yeah. So validity would be that the data is led based on business rules or calculation. The data is captured in a data store, can be through a graphical user interface, some background detail process with this data value according to the business rule. Yeah, Validities seems weird. Yeah, it describes the closeness of data values to predetermined values or a calculation. So is this data valid? How true is this data?

Linda:

Ohh well true. Yeah you can if you want to know how true the data is in the data warehouse you can again go to the source and if you really want to do. Kind of other things. Then you then you also check if the data of the source is really tuned, but that's too much for data in here.

Spreker 3 Yeah.

Remy:

Yeah, that's a bit too far.

Linda:

We must actually toss the data of the source and if we ever questions about it, then we can talk to the business like, yeah, I see this value here, but I don't think this is the value that should be here.

Remy:

Yeah, see another example of validity is for example birthdays menu system ask you to enter birthday in specific format and if you don't it's invalid, right? Yeah. So.

Spreker 3 Yeah.

Yeah.

Linda:

And that format should be also in the source system like input control like 18 slash. Yeah yeah yeah.

Remy:

Sorry, so it's. Set required. Or dash. Yeah. So yeah, the best way to deal with the validity is from the source itself. There's nothing.

Spreker 3

Yeah, so itself and.

Remy:

You can do.

Linda:

It's not, and it's not in the source. Then we should check OK what is in the source and can we transform it into the format? That's they. What they really want for each.

Remy:

What if the source isn't actually correct itself because it didn't have like? I bet it required field, you know.

Linda:

Yeah, if the source is not correct, yeah. Then we can try to make it correct, but if it's really the trash, then yeah, they need to update the source.

Remy:

Yeah. Yeah, fair enough. Yeah. This is also another thing that I just thought about because like, for example I right, you know, OK, the whole like most people places in Europe use for example, day, month, year, right.

Spreker 3 Yeah.

Yeah.

Remy:

Sometimes, for example me I I switch sometimes because spring is automatic right where I use month, day, year. But again, you wouldn't really know that I used month, day, year because let's say the month is 3.

Spreker 3

Yeah. Yeah. Yeah. Yeah.

Remy:

The days too, so I would think it's the 3rd of February. You know what, I.

Spreker 3 Yeah.

Yeah, yeah.

Remy:

Mean how do you deal with this? How do you even know?

Linda:

How do you deal with if the source is the other way of a year's time and?

Remy:

Or or it's a mixed sources of mix. Because I'm assuming the nurses that work with Axiom continue adding just old Dutch, right?

Spreker 3 Yeah.

Linda:

And ohh, but then you can write that case when statement actually case when date is this format then changing in this format and when in that format we also change it in this format yeah.

Remy:

It's forming. Yeah, that's interesting. That's very smart. Yeah. OK.

Linda:

And then you have all the. Then you see that everything will become one form. Yeah, that way. Yeah.

Remy:

Timeliness. So how do we measure timeliness, I assume?

Linda:

All that the data is on the right moment for the business. Ready.

Remy:

Yeah.

Linda:

Yeah. And that for that we have a job. Figures actually. So once in the night, like at 4:00 AM in the morning, 7:00 AM. And in the evening 7 then the the pipeline is right. Yeah, to be sure, that's multiple times that we really have the data. And then I think at my previous job, we had a troubleshooting. So at 8:00, we need to start work because at 9 the business works. So if the data wasn't ready yet, then you could.

Remy:

Yeah. Monitoring. Yeah, I had the same I do every morning, literally. Yeah, but also like part of the timeliness is, for example, imagine me as a as a healthcare worker. Yeah, I wanna check my schedule, but it's not up. To date, yeah. Think. Yeah. Yeah, what you propose is the best solution for this, but how do we measure it? Like, how do we display? You get what I mean. Like, if I'm gonna make data, if I'm gonna make a data quality dashboard and I wanna display the timeliness of Minecraft, for example.

Spreker 3 Right.

Yeah.

Linda:

Of how long it takes to get the data or or when it's updated. So lastly when it's lastly updated.

Remy:

Yeah. How long? What is the best practises of it? That's a. Very good question.

Linda:

The best practises for the day timeliness that you can check when it's refreshed for loss.

Remy:

Tenderness of deer. So when was last refreshed.

Linda:

Yeah, you can do it in Power BI also or you can create an Azure data factory your dashboard with OK these other sources. And this last update, this one hour ago. Yeah.
So it's accurate actually because it's updated the same.

Remy:

But some systems, from my own experience is that, for example, it's 1044 here, right?

Spreker 3 Yeah.

Remy:

But in surface systems it's 8:00 AM.

Linda:

You know what I mean is in another country.

Remy:

I don't know. For example, Azure Azure itself does this, right?

Spreker 3 Yeah,

yeah, yeah.

Remy:

Especially function apps and logic apps like I worked a lot. With. Them and usually like when I refresh it I refresh.

Spreker 3 Yeah.

Remy:

It at 10. Yeah, I checked the monitoring of the logic app or the function app itself. The timing is different. It's one hour back. Yeah, yeah.

Spreker 3 Yeah.

Linda:

Then, if you know that there are time differences, you can actually also change the time so it's the same.

Remy:

I see you mean like? Add plus one hour, OK.

Linda:

You can transfer the ETC. Now for example the date adds or the -1 hour or something. Yeah, yeah.

Remy:

Or add flamer. That's a good one. But just thinking if it's ethical you.

Spreker 3 Yeah.

Remy:

Know to do that. I believe so too, but that makes sense.

Linda:

If you want consistency, if you yeah, that the time is the same this plate everywhere.

Yeah, this use the time zone that is preferred here and for the business also. Spreker 3

Yeah.

Remy:

All right, now we talked about uniqueness. Job duplicate is your best friend.

Linda:

Now and also uniqueness, you have also always a primary key, so that also shows like this this thing, this person is registered and this role and primary key and.

Spreker 3

Yeah.

Remy:

What we do in case we have the same two primary matter 2 rows, same primary keys, different people, different names, I think that's not possible, shouldn't be possible.

Linda:

No, it shouldn't be possible. And if so, yeah. Then we must check out what is wrong. But it is technically. It shouldn't be. Yeah, it shouldn't be. I mean, every record when you have a new

employee, for example, it has, it creates a new record birth date and all the other thing has value. Also to see the uniqueness. But if you have the key with the same name and maybe it is the same name with a different person.

Remy:

A new number, yeah. Or at different values different. It's. Yeah, that's also a possibility same. Same a different person. Oh, that's a good one.

Spreker 3

Yeah. So there are many, many ways, yeah.

Remy:

That's slip you one.

Linda:

But I know that the primary key must be unique. That's also you also make constraints and you also make indexes on the prime.

Remy:

Yeah.

Linda:

The key.

Remy:

Yeah, that's fair. Yeah. Yeah, yeah, that's cool. Yeah, but not.

Linda:

So it's stage 1 so.

Remy:

Least. How do you measure the integrity?

Linda:

The integrity of the source.

Remy:

Of the server or just the?

Linda:

Data. Yeah, of the data you mentioned the integrity. Like I said, the the hashing, yeah, the hashing you create the you add a technical column which is called hash value.

Remy:

Oh. The hashing, yeah.

Linda:

And that that really we assured the uniqueness like every time a new record is inserted you have this hash value is all the time difference and it's based on the date the 2nd. So if you want to remain the history of the data.

Spreker 3 One

yeah.

Remy:

I see.

Linda:

Then you can see through all these hash values like. A lot.

Remy:

All right.

Linda:

Every day, every day, all the records get a new hash. Value. Based on the load dates and. Which is timestamp.

Remy:

I'm going to. Be very curious about the hash value, because it's the first time I heard about it, right? Right.

Spreker 3 Yeah.

Remy:

I think I might do my research on it early. If you could elaborate. More on the hash value itself.

Linda:

Yeah, the health value is a combination of all. All the fields in the record. So like.

Remy:

What do you mean by a combination of all the? Fields in the record.

Linda:

Yeah, the for example. You have one for one one row.

Remy:

Yeah.

Linda:

With primary key with the name, the source date then which is.

Remy:

Age, whatever.

Linda:

Name and other things. And then you actually concat.

Remy:

OK, you contact that.

Linda:

All the things together so you have so that's that's the long hash value. We have a lot of values actually.

Remy:

OK, so it's OK, the hash value is literally. It's the whole row.

Linda:

Yeah, a combination of the Oval, but in one field, yeah.

Remy:

Double row. Column. And if that is unique, then everything is unique.

Linda:

Yeah, that's that's unique.

Remy:

Yeah, that's main opening.

Linda:

And that's unique. Everyday it stains is it everyday change change doesn't change, but every day you got the new table and then you make the history. And then based on the hash value you know, OK, this this is unique because this day this second, this minute this data was located. Yeah. Yeah.

Remy:

Yeah, yeah. Yeah. We got this bunch of combination. You never thought about this. That's honestly smart. I love that.

Spreker 3

Yeah.

Spreker OK.

Remy:

Yeah. Then I have. We've been through all the dimensions here of data quality. We've talked about all seven of them. Yeah, I've talked about all the questions.

Linda:

Ohh great.

Remy:

I think I really don't have any more questions for you. Oh, yeah. No, I wanted to run this by you because the other day I was writing some best practises and etc. But see again, as I mentioned earlier, I still don't have the data. I'm not sure how, like I'm not very sure concretely about the project itself, right, like.

Spreker 3 Next.

Remy:

How detailed? For example, the best, best, good and? The. Bad practises should. Be you know, because for example the one I made is very on a on a high level and very general level, you know like it's just a protocol as we mentioned earlier, right, like what do you do in case of etc etc.

Spreker 3 Yeah.

Yeah, right.

Remy:

So for example, for data accuracy, I wrote, the best practise was to implement automated data validation cheques at various stages of the data processing accuracy, use standardised data format, coding system to minimise errors as we mentioned, right. Spreker 3 Yeah.

Remy:

Bad practise could be relying solely on manual data entry without validation. Yeah, well, increase the risk of cases. Yeah, that's this like this just for data accuracy. Is this detailed enough for the person to know what to do in case like, for example now? And I got a new

project I have to collect data since that somewhere etc. But I also part of my responsibility as a data engineers data quality.

Spreker 3 Yeah.

Dan. Yeah.

Remy:

So I gotta read this document. Do you think this is enough detail like?

Spreker 3 Yeah.

Linda:

I think it's it's definitely enough.

Remy:

I also gave examples if you could see like.

Linda:

And also something good for the data accuracy. There's always loving, you know, loving and loving. You can see this moment. The data is changed. So it's also good to look at the logging. If if you cannot find if you want to know, OK if there's something went wrong.
Spreker But.

Linda:

Or something in the data. You can always look at login. And then you can see this moment data was changed. This this moment. Yeah, modified actually, yeah.

Remy:

Modify that as basis. But that also depends on the source that already exists part of the programme that they had, right? If it actually picks the timestamp or when the thing has changed, because I don't think we could create that ourselves right now. Yeah, the loggings.
Spreker 3 Yeah.

Linda:

Logging. The loading is the loading of the data warehouse is in Azure data factory. You can just go to a monitoring dashboard and if you have a long one.

Remy:

So if someone changed the source, you get to know in Azure data factors. OK, no, no, no. OK, that's what.

Linda:

But not the short notes. No. So we all checked to ask them, like, is there any we have to make this?

Remy:

I thought yay.

Linda:

Strict rules with them, like if anything changes in the source you would like to know it because as well Venkat many times that someone changes something in the source and we didn't know. And then next day we had the whole list of troubles because the the source change. Yeah. And they're like, how come how come we're we're looking at ourselves.

Remy:

All troubleshoots.

Spreker 3 No.

Linda:

They tweeted something wrong. And the supplier, they changed the source.

Remy:

I see. I see.

Linda:

And also like you said API when they change the settings in the API and we don't know, yeah we can't access it. So it's all about communication also.

Remy:

All right. But like the example that I showed you is this, yeah, because this only shows within the dimensions of the data. Quality, but it doesn't really puts much of responsibilities, roles. What happens?

Spreker 3 Yeah.

Linda:

Man, while responsibility of the data supply that the data is available from their sites and

they. Spreker 3 And.

Linda:

Supply it to us so we can extract the data. So that's definitely the responsibility of the data engineer. Is the entire ETL process extracting, transform and load it and then the

responsibility of the BI specialist is to create those reports so the business can use the reports.

Remy:

Yeah.

Linda:

For KPIs or for the operations.

Remy:

All right. Yeah. So very general question. And to summarise the whole thing, yeah, so the question goes like how can a data quality metric be effectively designed to measure accuracy, reliability and serviceability of the intended purposes in healthcare data. So we mentioned is that the best ways to approach this is to actually follow the. Data quality framework, yeah, it mentions.

Spreker 3 Yeah.

Remy:

Stick closely with the business.

Linda:

Yeah, closely with the business, definitely like.

Remy:

Clear, clear, clear communications with the business. And that's the best way to approach this we also.

Linda:

And also to make a watching matrix so you know that this is the responsibility of this and this. Yeah. Yeah. That's also very important in the beginning of a project just to make sure, like, OK, we're doing this project with you, but we should make clear.

Remy:

Matrix. Yes, that's.

Linda:

Who's responsible, accountable for what and how many times you get back together back together to discuss the progress? And is there anything not clear? So it's communication very important. But also knowing what is my responsibility and accountability and stuff, yeah.

Remy:

And yeah, yeah, yeah. We also discussed about key components for automated consistency cheques that should be considered, so that's perfect.

Spreker 3 Yeah.

Yeah.

Remy:

And we also kind of discussed how robust data governance dashboard could be built.

Spreker 3 Yeah.

Remy:

I love this honestly, so I guess, yeah. We're done. Yeah, it was perfectly an hour information.

Spreker OK.

Linda:

Next interview will come next week or.

Remy:

I don't know because.

8.6. Appendices F: Interview Rink

Audiobestand rink

interview.mp3

Transcriptie

Remy

Well, good morning, rink. Yeah, thank you. Yeah. Could you tell me a little bit more about yourself, about your expertise, what since you have worked also in a project of data quality and consolidation of multiple sources into one?

Rink

Yeah, I think.

Remy

What can you tell me about what you've done exactly during the project?

Rink

Well, when I started in IT in 1998, I I I I dated this project.

Remy

MHM.

Rink

For a hospital. It mistakes the academic course premises within the the Pulmonary disease Department and they were migrating the from the old system as far as you can call it. Actually a system to a new system, which was a pretty full fledged information system. And but they had data in the old system and it needed to be migrated to the new system. Have you run into? Into a lot of issues in the sense that the data model of course is different. So so first you know you have to translate from the old data model to the new data model and well first thing is that all the records that cannot be translated for whatever reason. Yeah, you need to do something with it because you want to keep your history of your data in. Correct. So you have to decide on how to approach. Yeah. Everything that cannot fit in the new. Data model. And then the other thing that was a real pain in the neck were the business rules, because the old data was just put in the system without proper validation of the business rules, so.

Remy OK.

Rink

All data. Could could get in? Yeah, even if it was not correct or anything. So. And then the new system, there was a very strict appliance of of of very strict validation of data of the business rules.

Remy

To be anything and everything. Like a birthday requirement that it hasn't must be before or whatever after, yeah.

Rink

Yeah. Yeah, exactly. Yeah, yeah. So. And and these and these business rules, they were they were enforced upon the database with all kinds of insert triggers and update triggers. So there was like a shell around the database that was actually validating everything that that wanted to. I mean and in the migration that led to a lot of issues. So we had to build, we had to build a script that would let go that would let through everything that was OK.

Remy

MHM.

Rink

But everything was not OK you you didn't want the script to to crash and and and fail because then only part of the data will be transferred and the other part and you don't know exactly what you're missing. So we so we have to create this this script that will do first all sorts of validations if OK.

Remy

Well, what's going on? Yeah.

Rink

That's good. Could continue if lowercase it was written to A to a locking table, another table and and and it was used for further analysis. Yeah, based on the further analysis, we would enrich the script again. So we would actually, yeah, yeah, we would. We were actually analysing all the different types of exceptions that you could have. And then at some point we captured everything and we had one script which first went through all the except. Corrected them, inserted them and and.

Remy

Well, how do you? If you are, if you have corrected the data in the correct way, you know what I mean, especially that it was medical data, so you.

Spreker

Yeah.

Rink

Together with the people that were responsible from a functional point of view, so we were actually talking with the functional level ground level laboratory laboratory guys that were actually had had had functional knowledge of what was happening. Remy

OK. Yeah. Ah. Guys.

Rink

In the preliminary department, with all the research that they were doing with patients.

Remy

So mostly it was like back to manual work of validating manually with the laboratories.

Rink

Yeah. Yeah. So the so the. Correctness of data and the and whether or not the data should be migrated and based and and and how it should be mutated in order to comply with the business model and data model. There was all done in in conjunction with the people that knew functionally. What was going on? So this was, yeah, so, so of course me as an IT guy, I'm not gonna decide on on correctness or not. Correct. Yeah, so.

Remy

What is right word? Yeah. That that's interesting. So I see. I I see that you have worked through the whole project and you've worked through the whole data life cycle.

Rink

Right, yeah.

Remy

I see. I see. And alright.

Rink

And and and. Maybe one thing to add to this. This was a decentral information system only in this department, but you can imagine if you if you enter a hospital you are registered centrally in the central hospital system where your name, your Social Security number, your BS. Is being used to check your. Had your your the place where you live but also who is your General practitioner, but also where are you insured? Are you insured and how has your insurance been?

Remy

Yeah, I'm sure. Yeah. All the information about you basically, yeah.

Rink

Yeah, and this information is required in in all the systems in the hospital. So there need to be an exchange of information from the central system to the decentral systems. And I'm talking late 90s here and this was before HL 7 as a as a communication standard was actually yet defined in.

Remy

Yeah, it showed 7.

Rink

And that I'm being defined. And we also have, yeah, quite a. Were quite some issues with the exchange of data between central and D central because it needed to go both ways because we could also change. Things in the system in our department which needed to again be processed centrally as well as the. Other way around? Yeah. So it was bit and of course there was a delay because if if someone was centrally registered then it took like a day before they were also visible in the central system. But but the person comes in registers, walks, walks on to the pulmonary that is.

Remy

Yeah. But his data is not. There, yet yeah.

Rink

But the date is not there yet.

Remy

Yeah, I've seen similar situations like these, but wouldn't say it was in the health department, but mostly in Financial department. So you registered a new client and it appears next day in the database and like.

Rink

Yeah.

Remy

But there there's timeliness error in there. But you mentioned at Shell seven, health level 7 isn't health level 7. In general it goes more deeper into the data quality of within the health, but more detailed like blood pressure. Etcetera. Tumour size etc. Or is it? Just very general. Because compared to ISIL, for example.

Rink

It it it tries to capture all the all the information, all the elements of information in the hospital setting. It tries to capture everything.

Remy

I see. Would you consider like as I mentioned earlier, it during our first meeting like this is actually continue as a company for like how do you call it? Like forgot the wording for it. Old people's home. You know where? Where they provide care and etcetera. Is this like resulting yes. And is this like similar to actually being in a hospital data. You know what I mean? No.

Rink

Sorting style. No, it's not in the Netherlands. You have two types of of of care. You have cure and care.

Remy

OK. Yeah, I see. Yeah.

Rink

Here, here is hospital setting and also like a physiotherapist and stuff like this where you have an A problem that needs to be resolved. Yeah. Care is when you have a most for most times a chronic disease. And you need care, so you need someone to take care of you, to nourish you. And this this is this is also by law. It's differently arranged. So care cure is arranged in the servlet which was changed in 2006 for the worse by the way but the.

Remy

Took. Yeah, basically. Yeah. Ohh my God.

Rink

And care is used to be arranged in the alley bajet the ochman events be shown the visitor Coston.

Remy Yeah.

Rink

But they but it was like €40 billion per year went through the AVB awbs, yet it was centrally central money. It was from. The central government. And they they dismantled the the the TV set and they put a lot of it in the VMO the vet match on the stoning. The law societal support where they say if someone need to be taken care of, it can also be done by people in the municipality where someone lives like family or neighbours or.

Remy

Sports, yeah. Yeah.

Rink

Or the order means penalty must arrange something locally. Which seemed like a good idea, but it's also. It's also this has got a lot of disadvantages.

Remy

I see, but it is.

Rink

Especially when you talk about exchange of data, the more you, the more you you organise decentrally.

Remy

Yeah, yeah. Rink

The the Yeah, the the the more difficult it becomes to exchange the data properly. Remy

That's true, yes.

Rink

Better question for you because we talk about data quality.

Spreker Yes.

Rink

Do you consider different aspects of data quality?

Remy

Yeah. You mean like the dimensions of the data quality?

Rink

Yeah, like, how do you define quality of data in terms of which aspects can you look? At.

Remy

That is a very good question. Yes, like depending on which phase are we talking about within the the data life cycle, because within every data life cycle, there's specific ways to measure the data quality like. Which we come to to this like I have a lot of questions for the data quality. But yeah, how do you measure the quality itself? Well it it has to go behind it to hand with the business like you get it predefined it with the business like what is acceptable level of completeness. Yeah, what is an acceptable timely manner of like.

Rink

Yeah. OK.

Remy

When does the report arrive? Etcetera, but all of this, all the decisions here.

Rink

Yeah. OK. So so the so the things like completeness, yeah, correctness. Timeliness. Remy

Jumpiness, uniqueness, etcetera.

Rink

Continuity, that is, that is. Yeah, OK. This so you have you you are aware of the fact that you can break quality down and yeah, OK yeah. Notice this for me important to note that though yeah. OK cool.

Remy

In all the dimensions I've already. Made the research about it. So like we, we can also go through it, but also like to define these first, should we actually need to analyse the data or see what is complete, what is not and then you go to the business, what is an acceptable level?

Rink

Yeah.

Remy

Completeness. Yeah, that's my project yet haven't received the.

Rink

Exactly, yeah.

Remy

Data yet so I really can't do that analysis. But thank God. That the. Research question is part of the analysis. So which is? I think I've split it in a very good way. That first question is mostly focused on the research part, understand what is data quality, what goes around, et cetera. As an analysis. And then you could go to the business, etc. And then you could develop the protocols or whatever they want, right?

Rink

Yeah. Yeah, of course, yeah. And then use that as your.

Spreker Yeah.

Rink

Yeah, cool. Sounds, sounds sounds good.

Remy

I'm ready for. Please let everything alright. Alright. During the data collection phase.

Rink

OK. The data selection phase, yeah.

Remy

Collection fees, yeah, so as assumed. Like for example in the previous system or the old system you call the gas collected data from an API or some sorts. I don't know what was existing back then.

Rink

No, just there. It was just a database.

Remy

A database. OK, all right. All right.

Rink

And it was just a database, yeah.

Remy

And did you guys, uh considered any?

Rink

And I'm now, as we talked, because it came from a Clipper application. Clipper, you know.

Remy

OK. No.

Rink

It's on DOS, it's on Microsoft Dosh and that was it.

Remy

I don't know. Microsoft does, you know I don't have Internet.

Rink

You know, OK, well, why? It doesn't matter. But but they had a Clipper application, so the the Clipper, the, the, the they probably made some extraction from the data from Clipper and probably loaded it into a database in Oracle because we were working, we were rebuilding in Oracle technology. Yeah. And I reckon that they that they just exported the data from Clipper imported in Oracle in order for us, you know to to be in one environment where you can access the old data and migrate it to the new to the new tables. Yeah.

Remy

The. Data. But have you guys like for example used any techniques or tools to minimise data errors during? The entry level of the data or the transforming transforming of the data.

Rink

No, as I as I recall, it was just a dump of all the data.

Remy

Just I see, so kind of the same thing we're doing, which is because I don't know if you know the medallions levels.

Rink

That was there, yeah.

Remy

In the other bricks.

Rink

No. Yeah, you of course. You talked about about.

Remy

And so we have, yeah.

Rink

The silver and gold.

Remy

Rink

Like silver, gold and bronze. Yeah. So the bronze is dump everything OK. Well, bronze is you literally dump everything so and you don't care about anything. You know, just jumper.

Rink

But but I but I don't know. Exactly what it means. Yeah.

Spreker Yeah.

Remy

All right. And OK and have you, why have you guys stopped? Give it a thought on why did you guys choose to dump everything and not for example validate some stuff before you dump?

Rink

Yeah, that was the approach. Yeah, don't practising. No. Well, I was not involved in that part of the project, so I wouldn't know. So I just started in, in, in the environment where the old data was there and they needed to go to the new. But I know it was just a dump of everything because there was also a lot that that was so obsolete.

Remy

OK. All right. To mute it.

Rink

That we could actually get rid of but, but since there was an academic centre, there's there's also a thing where you're not an academic centre you you only need the information for your primary cure process, right? For your primary care process, the way you are in academic hospital, you also want to use this data for academic research.

Remy

Hmm.

Rink

So, so you so you require. More history, long longer term of your of your data and then then it needed to be anonymised and blah blah blah blah blah blah it was yeah. Remy

Linda comes out. Yeah, that, that, that I can imagine that's a lot of work regardless, yeah.

Yeah, but legally it's also it was also a thing because legally you are allowed or maybe even obliged to keep it for a certain amount of time. But after that you have to get rid of it. But you don't want to get rid of it because you want to need it. You want to look for a long term look to.

Remy

Rink

Make it anonymous.

Rink

That, but then you have to another.

Remy Yeah.

Rink

And when you're minimising the decentral system, you miss the the link with the central billing system for instance, and then you get this the these stupid errors between different systems.

Remy

That's. Ah yeah. The systems. Yeah. And then, honestly, I've I've never thought about that, but I can imagine it was like, really different and OK, OK, because you've said you are you you weren't really involved with in the data storage level either, right? Data processing level were you involved in the data processing level? I guess right. And have you taken any cleaning or pre process? Yeah, have. You. Done any cleanings for the data before? Like handling missing or inconsistent data processing.

Rink

Yeah, I was. Yeah. So, so, so in the process of analysing the data, sometimes people would say, oh, yeah, yeah, no, we know this goes wrong, but that's not an issue because the data again, we can get rid of anyway. Well, when when someone would say that, then we would actually delete it from the. From from from the source data, yeah, in the York environment. Yeah. Yeah. So we deleted it from the source, so we didn't think about it anymore. No. Or reflected. I think we actually flagged it like like we flagged the records as don't use them.

Remy

Hmm.

Rink

Because obviously. We didn't. I think we didn't delete anything, but just for. Remy

For for reasons we see saved it there.

Rink

For safety. Yes. Yeah, yeah. We just selected like and and neglect these records. Yeah. Yeah. And. And everything that we could not not neglect. Yeah. We we assess all the different exceptions that could occur.

Remy No.

Rink

Rink

Just as low as we as there were no exceptions to be found again.

Remy

Fair enough. Yeah. Yeah. But was it, if if I get this straight, it like the old system was only one system, one database, right. And it was it, like, separate different sources or something.

Rink

No, it was one database and I think it was I'm I'm trying to remember if it was a hierarchical. Or a relational database, but it could. It could well be a hierarchical database. Actually I think that was it.

Remy

Hmm. Alright.

Rink

Not entirely sure.

Remy

Yeah. OK. OK. I see. I see. Uh.

Rink

And there was. Also a big thing when you had. In your system, not all fields were. Mandatory. All the time.

Remy Yeah.

But in time some fields became mandatory, but then in your historical data you have some cells that are blank. And then as time progresses and these cells are being filled because they have been made mandatory in the in the application and in the migration, often we have records where some fields were empty that needed to be and they were mandatory also in the new.

Remy

Yeah. Rink

System. And we had to. To think about it, like which what data are you going to put in? Because you need to put in data because the database requires some data. But what? What will it be? Then you have to put in some dummy data. Yeah. And that people know that there's dummy because in some cases some columns that were mandatory, they also had

Rink

the uniqueness constraint. So you can fill it with dummy data, but it has to be a different. Yeah, a different measure every time or a different value every. Remy

OK. You need, yeah.

Rink

Time. So we had to run a sequence for.

Spreker Uh.

Rink

For those cells I was pretty annoying.

Remy

Yeah, I can imagine so. But I mean, but that is the way to ensure accuracy in your data to make it required fields with unique values.

Rink

Yeah. Or you just or you're just or you just inherit the mess that you have in your old days. You inherit it in your new database, and the question is do. You want that that.

Remy

Yeah. OK. Yeah. Is there a way to improve it if they want the data, but they don't know how to improve it.

Rink

I think the only way to improve this process would be to very to be very clear in the and asking the question what do you want to do with the data. If you want to keep it, why do you want to keep it? If it's only for reporting sake, maybe you can extract it in another way, put it in a different place where where where the rules are not that strict and use it. Only for reporting purposes or because of course you know you are, yeah. Polluting. Remy

The data. Yeah. Hey, I agree.

Rink

Your new database. Yeah. Is is. It's like taking The Dirty **** into a newly built house. And then. What? What? What is this?

Remy

What is it smell in here? Yeah, yeah, I know. I agree. I agree. But then at your expertise, what is the best way to ensure accuracy in the data?

Spreker Yeah.

Rink

Rink

This all boils down to to being to have very good, good conversations with people that are functionally responsible for everything that goes into the application. So people from the business, we call it an IT, they need to define.

Remy Yeah.

Rink

What is? What is good and what is bad? We will not define it as it, but we can be the critical people that, that, that that can push like I need the definition on what is good and. Remy

That.

Rink

What is bad?

Remy

Yeah.

Rink

The more you have before you start, the more easy it will be for you to clean your data and then move it through the rest of the of the of the process.

Remy

What's the? Here.

Rink

What we did was the trial and error rate so so so I would, I would just run the script and based on the new rules, I would look at what does. Comply. And and and and then we had the discussion. So it was very iterative, but it was also you could also call it trial and error.

Remy

What kind of metric did you use for the accuracy or the measurement of accuracy? If you remember. That's a good question in it.

Rink

Yeah, but I think we really took it from the from the, from the. From the old fashioned approach of day of data analysis and process analysis. So we so in the in the discussions with the customer with the pulmonary department. Which one are the rules that apply to the data model but also the business rules that apply in the process? And we, because this was the basis for the entire information system that we built and the entire data model that we. Designed. So. So we did all this pretty extensively in the system design phase for the new system based on the process that they that they wanted. So actually we did a lot of this.

Remy

Thank you. I hear you.

Rink

And and and then. Yeah. And then we and then the next check was, does the old data also comply with with this well with with the way we envision how the process should work and how the business rules should work and if not then per per case we would assess like how bad is this? Do we need it? How can we how can we cover this? Like either add data dummy data to the records or maybe remove or change data in the records. I mean the things that are that are just incorrect like wrong date format.

Spreker

Hmm.

Rink Or.

Remy

And I was going to come to that literally consistencies.

Rink

Yeah. Yeah, cause this. I mean those are pretty easy. Those can be also be analysed by by the IT people even.

Remy

Yeah, if if you have access to the previous source because you could compare, I think because I don't know for example, let's use the formats, right? I don't know if it was for example in the previous system an integer or a Boolean, right? But I know in the new system I see yes. No. Yeah, nay.

Rink

Yeah. Yeah, yeah.

Remy

So is this a billion or is just string? Yeah, right. No clue.

Rink

Yeah. Yeah, but we were looking only in the database itself, so if it was and and then and then in Oracle we it doesn't exist. So in so in Oracle it's either one or.

Spreker OK.

Remy

Ohh yes Sir. Zero, yeah.

Rink

If it's, if you only have two values and you want to keep. It as a. Number or actually what we did when it was, yes, no, we would just work with yes. Because it was. Performance wise, almost the same as one and zero. OK, that was that was not not a not a not a, not a big issue, no.

Remy

An.

Rink

No, no, right. So no things that could be an issue. I didn't encounter it in this project that I did encounter this in.

Remy OK.

Rink

A different project. When you set the the NLS settings, the national language settings, if you put them to. Yes. You get the month first, then the day, and then the year. Yeah. So 71 it's not 7 January. It's the 1st of July and and and and. And this is quite tricky. So you have to be aware when you work in the database that you have the proper settings.

Remy

Yeah. And they. Yeah, yeah. Also in the Netherlands, I've noticed this just the difference between English and Dutch which till today I'm so confused about.

Rink

Yeah.

Remy

So the commas and the points right? It confuses me like I know in English it's like, yeah, you use a point for a decimal. Yeah, in a coma for like if the. Yeah. And it's the complete opposite here. So the system you actually get to pay attention to the system. Yeah. Like hmm.

Rink

Yeah, it comes before, yeah. A decimal separated. Yeah. Yeah, yeah, yeah. For the for the thousands, yeah. Yeah. Yeah, yeah, yeah, yeah. And. And we also have this one nice example of where your data quality is impacted by the database technical database settings. We this was for Phillips Phillips lighting and they had a distribution centre in Spain and we got the we got an incident reported that when they print a sticker with an N with the thingy the wave. Yeah, I don't know the the the the English word for it but the same the symbol. That you have in Spain like espania, you know the end with the. Yeah, the accent. That looks like a wave.

Remy

And. Accent basically. This one right? Yeah.

Rink

That well. But every time they typed that particular character from the system, they got a cue. AQ like the the letter Q. So it would say espanca instead of Espana and and they couldn't find out what it was so I was sent there and I was looking in the database and in the database it was all.

Remy

A cube OK.

Rink

It it it I? Thought it was good, but it these were character based printers so you have to have to send the character string to the. To the to the to the printer. Now in the end it turned out that they configured the database or or the operating system with a 7 bit character set instead of the 8 bit character set. So if you look at the letter Q7 bit and you add the, the eight is the most.

Remy

And eight bit scanner I see. Yeah.

Rink

Significant bit and you add that for with. I think it's like 65 or so I think 65 you have to add to the ASCII code and then if you if you've tried you will get.

Spreker Right.

Rink

The end. Finally we found this but it it was it was very persistent nasty issue that was very difficult to find. But of course when you know.

Remy

You can imagine.

Rink

Yeah. After that is easy and then just funny, funny story one week or a couple of weeks later I was called. The printer is again not working. It doesn't work at all. You need to come here as soon as possible. That afternoon I actually went on a plane to Spain. I went there and I looked at I. I looked in the system. I couldn't find anything. Everything was just properly parsed from the database to a character based file that was sent to the printer. And then? In the printer, the printer will do some things, but there was no output and now it turned out that these printers they require this thermic paper. And they put on non term they put in non non termic paper so so there was no not like ink or anything.

Remy Yeah.

Rink

Yeah. So it was just a a **** ** of someone there.

Spreker

This is a technical issue of.

Remy

You guys, you know this.

Rink

Yeah, but it, yeah. But but. But for me, it was good because I was a recent pain again. Yeah.

Spreker

The 19th. Stupid. That's paid.

Rink

Yeah, yeah, yeah.

Remy

Oh, that's lovely. Yeah, yeah, yeah. For example. Like, we're going back to the consistencies topic.

Rink Yes.

Remy

Yeah. So I've seen like some parts of the data where it's like after the transformation after they've dumped everything to the bronze layer. So it comes to silver layer, we pick and we start doing data quality analysis, clean ETL processes, right. In general, now we've noticed that some formats caused.

Rink Yep.

Remy

Change in the values. So for example for example, as I mentioned earlier in the previous system, let's say it was an integer character 4, right? And then. But in the new system it shows. Imagine a house number right in the old system it's 24 House #24. In the new system it's 0024. Something went wrong in here, right? What is the best way to approach this? Mind you that all everything was dumped into. The new database so. So something must have went wrong there in. The data collection parts I assume.

Rink

I'm not sure if I. Have. Enough contacts. To to actually answer the question, but.

Remy OK.

Rink

When you move data around. Yeah. You want to you want to ensure that the data types are.

Remy

Still the same, yeah.

Rink

Are the same from from from the source? To the destination and and if you want to change it, which of course also happens. We had it, for instance, we work with the patient number, but the patient number was defined. I think 7 digits, but it had reached the end of the OR of the sequence.

Remy

Yeah. MHM. Characters, yeah.

Rink

So we need to go to 8 digits. Well, that's not a very complicated one, but also we have this code for the thing.

Spreker Yeah.

Rink

When you do an examination with the patient, you do several things and everything that you do is called affecting. I don't, I don't know like like some sort of transaction within the examination and they have a code and this code used to be it. It used to be numbers in the old system but it but they.

Remy

OK, I don't know either, but yes.

Rink

There were letters added to it in the new system, so it went from numeric to alphanumeric.

Remy

Alright, yeah.

Rink

UM. Which is I think also not that big of a deal, but what was big of a deal, what was a big deal was that these codes, they were unique. They were unique, but they were. But on. Oh, yeah. Yeah. For every for everything. There was a list of. Parameters that that were that were measured, the measurements that you did at a certain examination, you know, for instance, 1 predicting could be a flow with pharma cars. So you have to blow into a tube and you first do it without medication and then you do it with an education and in the tube.

Remy

MHM. Yeah, yeah, yeah, I know it.

Rink

The the the the the output of your lungs is. And there's all sorts of values that you can measure in the air that you breathe, and these were the parameters and there would be a normal value and the measured value and this would mean it's good or not.

Remy

Sure. Rink

Good. But the parameters that you use for each of these for each of these examinations? Uh, they could change in. Time, but the code of the examination itself remained the same, because this was used for billing at the insurance company. So so we had to make sure that something that was unique was unique within a certain time frame. So we had. So it was in the database, not unique, but it was unique.

Remy

Is the same.

Spreker

You see.

Remy

You can't. In this specific time.

Rink

In functional terms, it was unique, but in time it was not unique. So so this was also a nice a nice one, it's.

Spreker

I see.

Rink

With consistency, it's not consistency. Let's say on the data point level, but it was consistency.

Remy

Hmm.

Rink

In in how you look at it from a functional point of view or a technical point? Of view.

Remy

Well, that's very interesting. That's like a whole different way to look at consistency, like in a different perspective.

Rink

Yeah, because I mean the client can look at something and say, well, this is for me, this is a unique identifier for this is if, if if I if I look at students for me a unique identifier would be first name and last name most of the times it works you know.

Remy Yeah.

Rink

One class most of the times this is a unique set.

Remy

Sometimes it's not. Yeah, yeah.

Rink

No, sometimes it's not, but it's very, very exceptional. But but of course, in the system of. It's very common that we have multiple brown differences.

Spreker

Yeah. Yeah.

Remy

Literally, yeah.

Rink

Yeah. So what is unique technically, what's unique functionally, it's also a bit of the, but on the other on the consistency part. Yeah, it also depends of course the the amount of data types that you can work with a million numbers integers versus decimals is of course a big thing. If you if you use your decimals.

Remy

Where is it so the point?

Rink

Yeah. What was it? Yeah, of course. Yeah. So that's that's. I mean if you have. Preceding zeros like 0024 in your example, I wouldn't mind too much because they will fall off anyway and on the reporting side you can always take them off again, trim them at the.

Remy

Just take them off. Yeah, yeah. But is this the best practise to do that or do you have to deal with it in a different way or from a source way or from the blowing itself way you know?

Rink

What I mean, yeah, I know what you mean, but I find it difficult to to to as I would just say keep it the same as much as possible and. If. You change it. There must be a good reason for it from a functional point of view, preferably. Or from the technical point of view, when for instance, I don't know, you reach the end of the sequence or something like this. But yeah, that's that's.

Remy

Yeah, it's it's hard, boy.

Rink

Yeah. There's no, there's not. There's no. There's not, like, it's not that, like, I'm thinking of a certain best practise, no.

Remy

Yeah. No. Yeah, because I honestly think about because this is, there's no really right or wrong. You know, there's always. You should just ask the business, go with them. Is this actually true? Then we could actually maybe change it in the. Database itself, right? Yeah, I mean.

Rink

That's true for almost every question IT right?

Remy

That's true, and there be something automated level. I mean in machine learning it's different. You know it's filling the data in it because it's all different thinking perspective. This is not a machine learning part. So it's.

Rink

Yeah. Yeah.

Spreker

Not really easy.

Remy

Yeah, timeliness, yeah, timeliness. How do we ensure that?

Rink

Yeah.

Remy

Well, actually, see, I'm not sure that timeliness is part of will be part of the scope of my project because I will be working on like 2 parts of the patients and financial data for example, right. So.

Rink

Really.

Remy

It really doesn't matter to look at timeliness into patients data or right.

Spreker No.

Rink

I don't think it's a big it's a, it's a big issue because you you are working with data sets that.

Remy

Issue in here.

Spreker You.

Rink

More, more or less. Yeah. You're you're. You're not working on a on a production system with the transaction database that's being changed continuously.

Remy

Jason, no. Yeah.

Rink

So if you extract data from a database and then you're going. To work with it. Then time is not that big of an issue anymore.

Remy

Depending see if if, let's say's done with my part of these scopes and then I move to the other part which will be the scheduling and the roster. And I think yes, timeliness would be very important right? Because yeah, the nurse would like to know when and where. Right. I'll just go to my client.

Rink

Yeah.

Remy

But is there any any way or yeah do you have any best practises to approach timeliness or to ensure timeliness in the datas between different? Like I assume like from my literature study,

whatever from what I've read is that you could set it, set up a refresh early in the morning right and and ensure that you don't use any data before today, etc. You could even mark it like make it which database is refreshed, which database is actually live, which database is not.

Rink

Yeah, those are the things that you have to. Those are the things that you need to take if you, especially if you take from multiple sources. Yeah, you have to make sure that that the data set. So if you take data from different sources and you do it in a sequence, so you do it, you don't do it all at the same time because it might also be that some process.

Remy Yeah.

Rink

Some retrieval the collection process might run for a different time in one source because it's bigger than another source which is smaller. You can easily more easily or faster extract the data.

Remy Yeah.

Rink

Of course, yeah, you need to. Yeah. The moment you take a snapshot. So so you have to, you have to time stamp what whatever you're doing so that you can. What's the what's the word is the word for this the the the.

Remy

Time stamps here.

Rink

And. Conciliation, conciliation. Well, anyway, why you have to you you have to as a from a technical point of view, you have to ensure that all the data you're looking at actually relates to relates to each other. And it is from the same time. So yeah.

Remy

Describe it. Yeah. Yeah, yeah.

Rink

So time stamp, whatever you're doing.

Remy

Typing time stamps. It could be a very important part of this, but it honestly depends also on the data itself if the timestamp. Already? Or do we have to make a timestamp for it, you know?

Rink

Yeah, that's a that's a good one. Often there is a timestamp of the moment the records are being created.

Remy

Yeah. But assuming that there is no time stamp, what is the best way for us to create a time stamp like every time we collect the data? OK yeah.

Rink

Yeah, that. Yeah. Yeah, I will do that. Maybe even. If you if you retrieve the data from monstrous and you actually insert it like records in the database, you can actually add a timestamp of that particular injured in your database per record.

Spreker

In search.

Remy

Yeah, yeah.

Rink

But then it doesn't say again from a functional point of view, it doesn't say much. It only says when when it was inserted in your database.

Remy

Data. Yeah, and not in theirs, yeah.

Rink

And then, yes, but but their but their their their insurance is probably locked I mean

these days every record has a timestamp of of insert or or or update yeah. Remy

As the time stand.

Spreker Yeah.

Remy

I was reading about the intake integrity of the data the other day and how do you ensure integrity? I would love to run this idea by you right where? Well it it's part of uniqueness and integrity at the same time of the data. Have you heard about the hashing protocol?

Rink

Yeah. I know what hashing is, I haven't heard of. The hashing protocols will tell you.

Remy

Protocol protocol is something I added into it because I honestly loved the whole idea of it. I've never. Thought. About it, right. Yeah, it's like you would collect. Per row you would collect all the data, concatenate them into one column and that makes this specific row unique, right? And any count it up and check if there are any duplicates of.

Rink

Yeah. Yeah.

Remy

This. And this ensures, according to the the study that was reading, this ensures the integrity and uniqueness of the records that you have. What do you think of this? What is your own opinion about this?

Rink

It sounds it, it's it sounds fair, it sounds logical.

Remy

Are there any big UTS or something in it or?

Rink

Well, what, what what? I think is strange is but I'm very much reasoning from a relationship relational database here where every record has a primary key which is enforced in the primary key constraint. So so so every record in this particular tab.

Remy Yeah.

Rink

Or from a source table will have its own.

Remy

You need primary key.

Rink

Yeah, unique primary key so, so. So if you then concatenate all the values and put it in a row, of course it's unique because the primary key itself is already unique.

Spreker Yes.

Remy

I agree, I agree, but I've seen some databases that really do not have any primary keys. It's just it just dumps right? And it's like I'm always figuring out. How are you even connecting it in the semantic model?

Spreker OK.

Rink

It just dumps.

Remy

But yeah, you actually see that. So you don't know if there's anything unique in this record or not or assuming this business assuming it's a reselling business, right, where every week they sell something on the same market, right? I'm assuming that this week we sold 60, the week after we actually sold again 60. That we kept, we sold 62, right. If I look at these records, I honestly would think the 60 and 60 are duplicates. There's no time stamp. There's no unique. The only difference is, for example within a field description where it's there's something written you know, like ohh. Instead of Kettle he they bought something.

Rink

I have a date when it's sold.

Remy

No, it wasn't. If I remember, there wasn't no. So we had to. I have no idea how my boss did.

Rink

Yeah. OK.

Remy

It honestly because I was still working part time there, right? So. I don't know I. Will not pull up into the whole thing.

Rink

Well I, but I mean it's no, that's a fair thing, I mean. If someone if. Someone else creates a an SQL query to collect the data from a source and they leave out the the the the timestamp or the data or anything. Yeah, then it would be big puzzle for you and you just don't know.

Remy

At. Time stamp or something?

Spreker

Right. And.

Remy

Now I was doing a bad job because I was like and she went back to the API and I was checking again the sources and stuff and I was like going through because this might have to drop us to data. You need it right?

Rink

Yeah, but there is, there is really the the then you have to assume things or or assume it from the data that you know, but you won't know actually you you won't know. So then if if this will happen I will actually go back to the to the query that retrieved the the the original information to check is there any column that they've missed and that would help me.

Remy

Yeah, you never. 15 yeah.

Rink

To determine this.

Remy

But what do you think of the hashing scenario in this scenario I think. There is a good metric for measuring the integrity and uniqueness the records.

Rink

Well, I mean, it's to me it sounds like a message that that could be used. It's I think it's performance wise it's it's terrible. You know you're going to concatenate data that you already have again. So it's double the storage, a lot of processing on record level and then the analysis over.

Remy Yeah.

Rink

Let's say if you have 1,000,000 millions of records, it's quite something to determine uniqueness in so many records, especially when they are this long. So from a performance point of view, I would say.

Remy Yeah.

Rink

Try to avoid this. Try to avoid it, but make sure that you that you can determine uniqueness in another way from the source system that that will be.

Remy

I've also heard all those reading about indexing the primary keys. What is that used for exactly?

Spreker Yeah.

Rink

Normally a primary key is indexed anyway.

Remy

Right. Yeah. So what would you give it an extra index? Like I was so confused when I was reading this. I was like, huh?

Rink

I think I I think what you're trying to get at maybe maybe because this is a data modelling issue, what you're talking about and we actually already touched upon this a little bit before when you draught the data model, you're just talking with your client. You know you're talking about, OK, what are the things, what are the entities? That you want to save information about and this information about these entities you have, we call them attributes.

Remy Yeah.

Rink

And we put them in the data model. Now every entity on the level of how the customer perceives this world has a unique identifier. We, as it nerds, we immediately add a primary key sequence to a table in order to identify each. Record unique.

Spreker Yeah.

Rink

But that's not the unique identifier that that the average customer will use now these days. We live in a world where every almost everything has its code, like a Social Security number. We didn't have that.

Remy

Yeah.

Rink

Before the systems came, before the systems came, we would have a first name, maybe a second, third and fourth name, and then we would have a surname and. Remy

No number on the ID or something.

Rink

That. No normal, no ID. With that we will just have our name and our date, date and place of birth and this would wouldn't always be enough to you to identify one person. Of course it wasn't waterproof. It was a foolproof.

Remy

OK. Hey. Identify one person. Hmm.

Rink

And and that's why at some point when the systems took over, we needed something else. So we. Came up with. That by then it was called your Sophie. Number your social fiscal number.

Remy

Sophie.

Rink

And and these are phrases. If I call my car insurance company, they ask me what's your postal code?

Remy Yeah.

Rink

And your house number and I say. What does my postal code code and house number have anything to do with the way my car is insured? Because when I'm in my car, I'm never in this postal code or this mobile somewhere.

Remy

Yeah. Yeah.

Rink

Then I can easily find your information. So also here the code thinking you know the the indexing has has took over but this is not how how the real world actually works. Yeah. So in the if you call your General practitioner and you say I want to make an appointment.

Remy

Yeah, I see. Yeah. Where?

Rink

With the doctor, they ask for your for your name and your birth date.

Remy

Investigate. Yeah.

Rink

There, it actually still works along the old lines. They never ask you for your base. Then they'll say, what's your name? What's your birthday? Yeah, because that's for them. That's their unique identifier.

Remy

Yes and no. Even in hospitals till today.

Rink

In hospitals? Yeah, because they are looking at a very limited decentral scope of information, whereas you. Whereas if you would actually want to exchange information between hospitals that this will not work of course because there are many people who are maybe not in your case or in my case because our last names are not that common, but then at some point in time it will be an issue.

Remy

No. Yeah. And I can imagine.

Rink

Yeah. Yeah, so, so, so. So let's say back to your question, the unique identifier of an entity is often a combination. Of of of data points like first name, last name and the date of birth. But this is the actual unique identifier that you should use for any instance of an entity. Any records. But then yeah yeah, but the IT people always.

Remy

Snap. I see, yeah.

Rink

Add some sequence to it and then that's the thing that you will work with, because that's easily you. You need that but not necessary.

Remy

Unique, basically, yeah.

Rink

True, and this is. Where your so also here you should if you run into. Let's say strange. Duplicates. You should also talk with your client, like how would you uniquely identify every instance, every occurrence of an entity that we store information about.

Remy

Jessica. No. Yeah. Yeah, that's true. That's true. Never thought about this in, for example, in uniqueness. It's like I never thought about asking the business because, you know, Python, one Python code will check per row the whole everything, right. Rink No.

Remy

Yeah, you could just do that. What would you ask the business? But it's honestly interesting to ask the business about this. What do they identify as unique? Yeah, that's good. All right.

Spreker Yeah.

Rink

Yeah, now if you. Look at cars, for instance. Cars you would say, OK, the so the brand of the car and then the type of the car and then maybe the let's say the, the, the, the specific stuff like this or but but also this is code fight Now because now you have the Chelsea.

Remy

Cylinders.

Rink

For the the what doesn't work for the? Very, very, very, very, very.

Remy

You know, I know venom bars and things. Yeah. Yeah, and yeah.

Rink

Very NVN number V number. And then this, if you look at this code it it contains everything of of that particular that particular unique action plan that particular. Remy

The person? Your car. I'm a car person II fanatic, I would say.

Spreker

OK. Yeah.

Rink

Yeah, but is this the naming convention for the VPN code like the first bit is the brand, and then the second is the type and then blah blah.

Remy

Same with. Tyres. The same goes for tyre size with etc.

Rink

Yeah, yeah. Yeah. Yeah, yeah. Yeah, yeah, yeah, yeah.

Remy

You would see it as just one number, but yeah, first three numbers tells you decide.

Rink

Well, fortunately, in the normal world it will be very common to uniquely code stuff. Serial numbers, VN codes, basens Social Security numbers. So that makes the life.

Remy

Yeah. OK. Rink

Of the IT people a bit more because you can just take that as a as a as unique, that's fair. But The funny thing is, umm, I hardly ever see a.

Spreker Easier.

Rink

Information about persons, whether they're students or employees, where the Social Security number, the basin number, is actually the primary key column, and.

Remy

Yeah, I think that goes to security reasons, right?

Rink

But I mean if the base is in, is a unique number identifier for four people. Remy

Yeah, I don't use that.

Rink

Why not use that as the primary key and then people say yeah, but performance and sequence. But yeah, yeah, it might be, but then you still need to put a uniqueness constraint of on the column with the base and and you still will have the processing to ensure uniqueness.

Remy

Thank you.

Spreker Yeah.

Remy

Let's see, yeah. Yeah, let's see from the processing part. Honestly, both will be the same trouble, right? So it's just about I think it goes to privacy part because I was, I was dropping some parts into privacy. What I'm reading on governments and because governance is not really part of the project because it's very.

Rink

There could be.

Spreker Sure.

Remy

Like I'm trying to limit the scope, but still I'm curious person. Let me read. Right. So I'll really.

Rink

Yeah, yeah, yeah, yeah.

Remy

It's really part of privacy where PS ends and stuff, even though they collect it, you're not allowed to keep it for like more than a year or something like that. And it's it's a whole headache. I was like, good luck.

Rink

Yeah, stuff like that. Yeah. Yeah, yeah, yeah, yeah, yeah. Yeah, yeah, yeah. OK. Remy

That's why I'm not going. Into privacy.

Rink

No, no. Well, it's an interesting field.

Remy

And then security and privacy like, yeah, well, we finished this. Well, I I guess I've got, well, everything from you. Thank you for your time. You're.

Rink

Well, maybe one thing that I could ask. When it comes to her, let that let the business

define uniqueness the way the business defines certain things that you look at. Spreker

Yes.

Rink

Can also be a bit of pain in the ***. I had this one situation at NXP once where we were making all the all the all the business intelligence reports for for the for the for yeah for corporate and and we would extract data from separate business units.

Spreker What?

Remy

MHM.

Rink

And we would report on, for instance, the amount of open orders. So it sounds fairly easy, like you extract from the order table and then you count every order.

Remy And.

Rink

It has status open or the status other than closed the bit depending on the question. But The funny thing is, is that what they consider open in one business unit is completely different defined than what they consider open in another business unit. SO1 business unit, as soon as the as the order was handed in, it would start as an open order and when it was delivered, it would be closed, whereas the other will say now.

Remy Yeah.

Rink

This this uh order is new and. Then it would. Be. Inserted or something like that and then it would be open and then. But then like you miss a lot of orders that you don't count. If you already looked at status open.

Remy

Yeah. Yeah.

Rink

So. We had actually a Bureau, a group of, let's say, four people that were maintaining the standards and definitions that were used within within corporate.

Remy

I see the consistencies between the corporates.

Rink

So. Yeah, but now you're in a very functional level, but the look can go wrong there on the IT part because, you know, we are just. Yeah.

Remy

So. Yeah.

Rink

Plain and stupid. If someone says count open we look steads open count. Yeah. Mean yeah, that's what you asked for. Yeah.

Remy

Sure counts open. Here you go. Just literally.

Spreker But.

Rink

Yes, so, but if you have multiple sources and and you look at these kind of statuses, make sure that you check is the status in line with the status, the with the definition that. Remy

Yeah. No, no, I get it.

Rink

You are using.

Remy

Yeah. No, that was also part of my idea, where as soon as I get the data, I'm going to do a whole bunch of analysis about everything before I go to the business. It's like, yeah. Rink No.

Remy

I think you should define this and that. This way based on this and that. But how do you want? To define it. Yeah, it's like you must gotta give them everything, right? Yeah. Yeah. Spreker Yeah.

Rink

So I think in summary, it's all about really understanding. The data that you're looking at and and and everything that you that you don't understand or yeah, don't assume, but make sure you validate with someone from the business or maybe even a couple because they often also don't know.

Remy

Itself. Hmm. I think a lot of also machine learning methods could be used in this some of them.

Rink

Yeah.

Remy

Like. Yeah, yeah. No, that was lovely. OK. I had a specialisation semester for an AI. I really enjoyed it.

Rink

Maybe. Yeah. Yeah. What did you do as Mr 7 then?

Remy

Advanced business and yes, Mr Seven was a minor.

Rink

Yeah, yeah, so 6. You did the vote, but yeah. Yeah, yeah.

Remy

Questions. OK. All right. But thank you for your time.

8.7. Appendices G: Advised protocol

DATA INTEGRITY ADVISED PROTOCOL

Remy Alashabi

8 APRIL 2024
TWENTYNEXT
Eindhoven

Inhoud

1. Introduction 144
2. Data Integrity definitions and criteria 144
3. Roles and Responsibilities 146
4. Data Collection 147
5. Data Validation and Quality Assurance 148
    5.1. Data integrity Audits 148
    5.2. Types of data to audit and not to audit 148
    5.3. Data Cleansing and Standardization 149
6. Data integrity Monitoring 149
7. Training and Education 150
8. Continuous Improvement 150
9. Business rules and metrics: 151

Bibliografie 159

Protocol for Ensuring Data Integrity

Introduction

Data integrity means making sure information is accurate, consistent, unique, and complete (Business insights blog, 2021) . In healthcare, getting this right is crucial for providing good care. This protocol sets out a plan for AxionContinu to maintain data integrity. By following these steps to ensure data is complete, accurate, consistent, and unique, nursing homes can make sure their information can be trusted and used effectively. This helps in making informed decisions, following regulations, and ultimately, improving patient well-being.

Data Integrity definitions and criteria

Data integrity has 4 dimensions, which are mentioned above. Each dimension has its own acceptance criteria and they are as follows:

Dimension	Definitions	Criteria
Completeness	Complete patients data includes all necessary information to effectively assess, plan, and deliver care (Anwar, 2024).	• All required data fields are filled out. • There are no missing values in crucial data fields. • Data captures the entirety of relevant information for its intended purpose. • Data collection methods ensure that all necessary data points are obtained. • Data existing in platform A must exist in platform B.
Consistency	Consistent data maintains uniformity and coherence across different sources and over time (Democracy westlancs gov uk, 2012).	• Data values are uniform and follow the same format throughout the dataset. • There are no conflicting or contradictory data entries. • Data follows predefined standards or rules, such as date formats or naming conventions. • Cross-referencing data sources yields consistent results.

Uniqueness	Unique data ensures that each patient's information is distinct and identifiable, avoiding duplication or ambiguity (Anwar, 2024).	• Each data record is unique and not duplicated within the dataset. Primary keys • or identifiers are used to ensure uniqueness across records. • Duplicate detection mechanisms are in place to identify and resolve any duplicate entries. • Data integration processes maintain the uniqueness of records when merging datasets
Accuracy	Accurate data reflects the true state of a patient's health status, treatment plan, and care interventions (Anwar, 2024).	• Data should be free from errors, inconsistencies, or inaccuracies. • Verification processes should be in place to validate data accuracy against trusted sources or standards. • Data collection methods should minimize errors during entry, transmission, or processing. • There should be mechanisms to identify and correct inaccuracies promptly.

Roles and Responsibilities

- **Data Stewards:** Data Stewards typically manage specific data sets within the organization, ensuring accuracy, completeness, and consistency while overseeing adherence to data integrity standards, including content, context, and associated business rules, while also advising on data governance practices and potentially sharing responsibilities with data engineers (linkedin, 2024).
- **Data Owners:** Data Owners are responsible for the data in their specific areas, ensuring accurate documentation and maintenance to facilitate safe and effective care delivery. They control and manage the quality of datasets, setting data quality requirements, and typically represent the business side at a senior level within the team (Alexsoft, 2019).
- **Data Providers (Healthcare Providers):** Healthcare Providers serve as the primary source of data input, responsible for documenting patient information accurately and consistently (linkedin, 2024).
- **Data Engineers:** Data Engineers manage the technical aspects of data integrity, including configuring systems, implementing validation rules, and ensuring system reliability and security during the ETL process (Alexsoft, 2019).
- **Data Consumers:** Data Consumers utilize data for various purposes, such as decisionmaking or analysis. They are usually the ones that define the standardized data and report errors back.
- **Data Administrators**: Data Administrator, also referred to as a Data Analyst, is responsible for processing data, deciding what's relevant for database storage, and overseeing data management. Primarily, this role is business-oriented as they ensure data integrity by allocating resources, setting policies, and ensuring compliance (GfG., 2023).

Data Collection

- **Standardized Forms:** Nursing homes should develop standardized templates and forms for capturing resident information. These forms should include all essential data fields required for comprehensive assessment and care planning. For example, an admission assessment form may include sections for demographic information, medical history, allergies, and care preferences (Linkedin, 2023).
- **Training Programs:** Staff members should receive training on proper data collection techniques and the importance of completeness, accuracy, and consistency in

documentation. Training programs may include hands-on workshops, online modules, and ongoing education sessions to reinforce best practices (Anwar, 2024).

- **Validation Checks:** Electronic health record (EHR) systems should be configured with validation checks to enforce data integrity at the point of data entry. Validation rules can include range checks, format checks, and logic checks to ensure that entered data meets predefined criteria (Anwar, 2024).

Data Validation and Quality Assurance

This chapter outlines internal processes and methodologies for validating data accuracy and maintaining high-quality standards throughout the data analysis lifecycle.

Data integrity Audits

Data integrity audits are conducted internally to assess the accuracy, completeness, and consistency of datasets used for analysis (Karl., 2024). These audits are based on IBM (Sluzki, 2023)and they involve the following:

- **Statistical Analysis:** Performing statistical analysis to identify outliers, anomalies, and data discrepancies. Data engineers utilize descriptive statistics, hypothesis testing, and outlier detection algorithms to assess data quality and detect potential issues.
- **Data Profiling:** Performing data profiling to understand the structure, patterns, and anomalies within datasets. This involves examining data distributions, identifying outliers, and detecting missing or duplicate values.
- **Error Identification:** Implementing algorithms and rules to automatically detect errors and inconsistencies within datasets. This may include validation checks for data range, format, and referential integrity to flag anomalous data points.
- **Root Cause Analysis:** Conducting root cause analysis to identify the underlying reasons for data quality issues. This involves tracing errors back to their source, whether it be data collection processes, data entry errors, or system limitations.

Types of data to audit and not to audit

Data integrity audits should encompass categories of essential data for the organization's operation and that includes the following:

Data to audit:

- **Scheduling data (mijncaress):** Data entities such as the activity of the employee, the care they provide, when and where do they have to be and etc.
- **Transactional/Financial data (visma.net):** Any records documenting business transactions, such as invoices, insurance transactions and etc.
- **Personeel data (afas):** any records about the active employees, such as their salary or their functions.

- **Sensitive data (mijncaress):** Any confidential or regulated data such as BSN numbers, health records, and personally identifiable information.

Data to not audit:

- **Temporary data:** data that is generated for a short time and not integral to any operations or tables that are not in use anymore.
- **Archived data:** Data that is no longer used for decision-making but is kept for compliance or reference purposes.
- **Recreational data:** Data that is not related to any business decision making decisions, such as employees vacation hours or date of birth.

Data Cleansing and Standardization

Data cleansing and standardization processes are employed to rectify errors, remove duplicates, and ensure consistency within datasets. These processes involve:

- **Data Cleaning scripts:** Utilizing algorithms and scripts to clean and transform raw data. This may include techniques such as imputation for missing values, deduplication to remove redundant records, and normalization to standardize data formats.
- **Standardization Rules:** Establishing standardization rules to harmonize data across disparate sources. This may involve mapping data to common vocabularies, resolving inconsistencies in naming conventions, and enforcing standardized formats for dates, addresses, and codes.

Data integrity Monitoring

Continuous data quality monitoring processes are established to proactively identify and address data quality issues in real-time (LakeFS., 2024). These processes involve:

- **Automated Alerts:** Implementing automated alerting mechanisms to notify stakeholders of data quality issues as they arise. This may include thresholdbased alerts for data anomalies, trend-based alerts for deviations from expected patterns, and anomaly detection algorithms for outlier detection.

- **Dashboard Reporting:** Developing interactive dashboards to visualize key data quality metrics and trends. Dashboards provide stakeholders with real-time insights into data quality performance, allowing them to monitor KPIs and track progress over time.
- **Root Cause Analysis:** Conducting ongoing root cause analysis to investigate the underlying factors contributing to data quality issues. This may involve collaborating with data owners, subject matter experts, and IT teams to identify systemic issues and implement corrective actions.

Training and Education

- **Comprehensive Training Programs:** Nursing homes should develop comprehensive training programs for staff members on data integrity principles and best practices. Training may include classroom instruction, hands-on exercises, and scenario-based simulations to reinforce learning (Anwar, 2024).
- **Ongoing Education:** Ongoing education and support should be provided to staff members to ensure they stay up-to-date on data integrity practices. This may include regular refresher courses, lunch-and-learn sessions, and access to online resources and training materials (Anwar, 2024).
- **Specialized Training for IT Personnel:** IT personnel should receive specialized training on implementing technical solutions to enforce data integrity measures. This may include training on configuring EHR systems, developing validation rules, and troubleshooting data integrity issues (Anwar, 2024).

Continuous Improvement

Continuous improvement initiatives are integral to maintaining data quality standards and driving organizational excellence (Foot, 2022). These initiatives involve:

- **Feedback Loops:** Establishing feedback loops to gather input from stakeholders on data quality issues and improvement opportunities. This may include feedback through surveys, focus groups, and regular meetings to identify pain points and prioritize enhancement efforts.
- **Root Cause Analysis:** Iteratively conducting root cause analysis to address recurring data quality issues and prevent future occurrences. This involves implementing corrective actions, revising processes, and enhancing data governance practices to mitigate risks and improve overall data quality.

- **Performance Metrics:** Defining performance metrics and benchmarks to measure the effectiveness of data quality initiatives. This may include tracking KPIs such as data completeness rates, error resolution times, and customer satisfaction scores to evaluate progress and drive accountability.

Business rules and metrics:

Based on the research done above and the guidance meetings with the stakeholder, the business rules were made and the metrics are also chosen based on them. The following metrics below are taken from 3 different sources and they are (Mssaperla., 2024), (Aliaga, 2023) and (Databricks, 2024).

A. Completeness:

i. Business:

- Check required fields: Specify which columns are important for decision making.
- Analyze completeness % on the columns individually, then per couple of columns combined.
- Cross-referencing: maybe cross reference afas with mijncaress.
- Compare the number of rows to the original source
- Check hoofd letter
- Check if the name once prof or professor.

ii. Metrics:

- Percentage of Missing Values: (Number of missing values / Total number of values) * 100
- Completeness Score: (Number of filled fields / Total fields) * 100
- Percentage of Records with Complete Information: (Number of records with no missing values / Total number of records) * 100
- Timeliness of Data Entry: Measure the time taken to fill missing data.

- Percentage of Mandatory Fields Filled: (Number of mandatory fields filled / Total number of mandatory fields) * 100

B. Accuracy:

i. Business

- Manual inspection: review each row manually, or random sampling(random select of random rows ex row 1, 20,100 etc) and look for the following :
    a. Data format: check if data format of the columns is as expected(dates are dates ints are ints and etc)
    b. Data range: check that the numerical values are within expected range (Dob can't be in
    1890 or negative number)
    c. Data consistency: ensure that the row is logically consistent (I.e., ensure that the gender matches the name)
- Cross referencing : compare the patents data with another data set, compare locations on google maps if they actually exists
- Check the input time that it is up to date.

ii. Metrics

- Error Rate: (Number of errors / Total records) * 100
- Consistency Check: Cross-reference data against trusted external sources.
- Data quarantine: Bad records can be identified and inserted into a quarantine table.
- Flag violation: Bad records can be added into the dataset with a tag that it needs to be checked.

C. Uniqueness:

i. Business

- Check for duplicates based on the required fields.

- Review the Key identifiers: Identify for each row the unique ID or as Linda mentioned, use the Hashing method, where I combine all the fields together per row to check for uniqueness
- Manual inspection per row: look for patterns as repeated sequences or unexpected duplicates.

ii. Metrics

- Duplicate Record Count: Count the number of duplicated records.
- Unique Values Count: Count the number of unique values in each field.
- Key Field Duplication Check: Ensure primary key fields remain unique across all records.
- Cardinality Check: Assess the uniqueness of values in a field.
- Cross-System Uniqueness Check: Ensure uniqueness across different systems or databases.

D. Consistency:

i. Business

- Check field relationship: in the same dataset, within each row check the different field s i.e patients information verify that the age matches the DOB
- Standardized format: ensure that data is consistently formatted across all the rows i.e (ddmm-yyy in all the rows and not suddenly mm/dd/yyyy) uppercase and telephone 06 number or 0031 and etc
- Uppercase and low case format

ii. Metrics

- Data Validation Rules Compliance: Measure the percentage of data meeting predefined validation rules.
- Referential Integrity Check: Ensure consistency in related datasets, e.g., foreign key constraints.
- Temporal Consistency: Check for consistency of data over time.
- Logical Consistency: Ensure data adheres to logical rules and constraints.
- Format Consistency: Check if data formats (e.g., date, currency) are consistent across records.

It's important to tailor these metrics according to the specific requirements of the specific project and characteristics of the dataset and business needs.

10. Mijncaress column definition

Table name	Kolomnaam -	Specifications	Note
TBC	IC	Primary key must always be unique per row. Can contain letter in case of alphanumeric.	
	LongNm	-	
	Nr	Range:1-5 Null count: 0 Count: 33334 Min: Max: UCase:	
	VrNm	Range: 1-50 Null count: 580 Count: 33334 Min:- Max:- UCase:True	
	CallNm	-	

	Vrv	Range: 1-10 Null count: 21130 Count: 33334 Min:- Max:- UCase: True Distinct count: 101	1 row looks incomplete and included a part of last name (an dr wal-)
	AchtNm	Range:1-40 Null count: 0 Count: 33334 Min:- Max:- UCase: True Distinct count: 12275	1 row with 40 length has other value than name (MIC Gevaarlijke situatie VH Vredenbur)
	Geslacht	Range: 1 Null count: 0 Count: 33334 Min:- Max:- UCase: True	Letters only

		Distinct count: 2	
	GebDat	Range: 3-10 Null count: 17 Count: 33334 Min:1-1-1900 Max:8-11-2022 UCase: - Distinct count: 15664	1 record where OvalDat was before the Geb dat. Length of 3 is nan

BirthPlace	Range: 1-35 Null count: 777 Count: 33334 Min:- Max:- UCase: True Distinct count: 2963 Pattern: ? count:34 Pattern 0, count 124 Pattern , 122 Pattern ., count 96 Pattern x, count 29 Pattern -, count 5 Pattern 0Fes, count 7 Pattern 1, count 2 Pattern 0Mut, count :2 Pattern 0Bor, count 2 Pattern ??Zmit, count 1 Pattern 14-02-1935 , count 1	1 record had 35 length (Ms “Sibajak” 47.57 N.Br en 6.14 W.L)

OvlDat	Range:3-10 Null count: 13203 Count: 33334 Min:27-9-1921 Max:16-3-2024 UCase: - Distinct count: 6410	1 record where OvalDat was before the Geb dat. Length of 3 is nan
Iusr	-	
Icountry	-	
Ilang	-	
Tel1	Range:1-24 Null count: 4080 Count: 33334 Min: Max: UCase: - Distinct count: 24956	

		Pattern: -, Count: 1870 Pattern: geen, Count: 3 Pattern: cp, Count: 35 Pattern: !, Count: 22 Pattern: zoon, Count: 4 Pattern: CP, Count: 21 Pattern: doch, Count: 8 Pattern: (, Count: 16 Pattern: ., Count: 4 Pattern: echt, Count: 2 Pattern: zelf, Count: 2 Pattern: 03, Count: 19260 Pattern: 06, Count: 6297 Pattern: , Count: 1960 Pattern: 31, Count: 480 Pattern: 01, Count: 330 Pattern: 00, Count: 278 Pattern: 02, Count: 273 Pattern: 07, Count: 104 Pattern: 05, Count: 68 Pattern: 0, Count: 65 Pattern: 0032, Count: 3 Pattern: 0035, Count: 2	What do these mean cp, echt, zelf and etc?
	Sofi_nr	-	
	BSN	Range:1-9 Null count: 582 Count: 33334 Min: Max: UCase: - Distinct count: 32730	BSN 0, count : 14 BSN: 111222333 count: 8 BSN: 00000000, Count: 3. BSN:099430022 Count:2 BSN 81745382, Length 8.
	Straat		
	Hnr		

TBC_AD	HnrToe		
	Pc		., 000, 9999vv 3632ES Loenena/d vecht doesnt exists it is EP
	Woonpl		

1. **Formatting Rules**:
    - Telephone numbers should follow the E.164 format, which includes:
    - Maximum length of 15 digits.
    - Country code as the first part of the number.
    - Remove any spaces, dashes, or parentheses from the numbers.
    - Ensure the correct international prefix is used (e.g., +31 for the Netherlands).
    - Geographic numbers should have a total of 10 digits (including the leading 0 when dialing within the country).
    - Non-geographic numbers may vary in length but typically adhere to standard formats specified for specific services.
2. **Geographical Numbers**:
    - Consist of nine digits.
    - Comprise an area code (two or three digits) and a subscriber number (six or seven digits).
    - Dialing within the country requires prefixing the number with '0'.
3. **Non-Geographical Numbers**:
        - Vary in length but generally kept as short as possible.
        - Mobile telephone numbers always have 10 digits.
        - Follow specific formats based on their designated use:
        - 06: Mobile telephone operators
        - 0800: Free service numbers
        - 084, 085: Used for VoIP

- 087: Voicemail and virtual private numbers
- 088: Large companies with multiple addresses
- 0970: Machine-to-machine communication (8-11 digits long)
- 0979: Machine-to-machine communication (no fixed length, reserved for internal network usage)
- 0900: Paid information services
- 0906: Adult lines • 0909: Entertainment

4. **Data Validation**:
    - Validate telephone numbers to ensure they conform to the E.164 format.
    - Check for the presence of the correct country code and prefix.
    - Verify that numbers are correctly categorized based on their usage (geographical or non-geographical).
    - Implement checks for common patterns associated with scam numbers (e.g., 066, 084, 087).
    - Ensure consistency and accuracy in storing and processing telephone numbers to prevent errors in communication or service delivery.

Bibliografie

Alexsoft. (2019, 10 17). *Data Quality Management: Roles, processes, tools.* . Opgehaald van AltexSoft: https://www.altexsoft.com/blog/data-quality-management-andtools/

Aliaga, A. (2023, 12 20). *introduction to Databricks Lakehouse monitoring - Antonio Aliaga - Medium.* Opgehaald van Medium. : https://medium.com/@antaliagacortes/introduction-to-databrickslakehousemonitoring-aebeddf013b5

Anwar, M. (2024, 4 4). *5 Crucial Best practices for ensuring data quality in healthcare. Astera.* . Opgehaald van Astera: https://www.astera.com/type/blog/managingdata-quality-inhealthcare/

Business insights blog. (2021, 02 4). *What is data integrity and why does it matter?* Opgehaald van Harvard business school: https://online.hbs.edu/blog/post/whatisdata-integrity

Databricks. (2024, 3 4). *Monitor metric tables.* Opgehaald van Databricks on AWS. : https://docs.databricks.com/en/lakehouse-monitoring/monitor-output.html

Democracy westlancs gov uk. (2012, august). *DATA QUALITY PROTOCOL*. Opgehaald van https://democracy.westlancs.gov.uk/Data/Audit%20&%20Governance%20Committee/201209251900/Agenda/023972_DQPROTOCOL.pdf

Foot, C. (2022, 5 6). *8 proactive steps to improve data quality.* . Opgehaald van Data Management. : https://www.techtarget.com/searchdatamanagement/feature/Proactivepractices-fordata-quality-improvement

GfG. (2023, 5 17). *Difference between Data Administrator (DA) and Database Administrator (DBA).* . Opgehaald van GeeksforGeeks: https://www.geeksforgeeks.org/differencebetween-data-administrator-da-anddatabase-administrator-dba/

Karl. (2024, 2 9). *Audit Analytics: types, benefits and use cases.* . Opgehaald van Caseware Canada. : https://www.caseware.com/ca/resources/blog/auditanalytics-typesbenefits-and-use-cases/

LakeFS. (2024, 3 11). *Data Quality Monitoring: key Metrics, techniques & Benefits.* . Opgehaald van Git For Data - lakeFS.: https://lakefs.io/data-quality/dataqualitymonitoring/

Linkedin. (2023, 10 5). *How to Improve Data Quality with Effective Data Governance.* . Opgehaald van Analytics8, Data & Analytics Consultancy. : https://www.linkedin.com/pulse/how-improve-data-quality-effectivegovernanceanalytics8

linkedin. (2024, 02 18). *How can you define data quality roles and responsibilities?* Opgehaald van linkedin: https://www.linkedin.com/advice/0/how-can-youdefinedata-quality-roles-responsibilities-vswge

Mssaperla. (2024, 4 3). *Monitor metric tables - Azure Databricks.* . Opgehaald van Microsoft Learn.: https://learn.microsoft.com/engb/azure/databricks/lakehousemonitoring/monitor-output

Sluzki, N. (2023, 8 30). *8 Data quality Monitoring Techniques & Metrics to watch.* . Opgehaald van IBM Blog.: https://www.ibm.com/blog/8-data-quality-monitoringtechniques/

8.8. Appendices H: Jupyter notebook screenshot

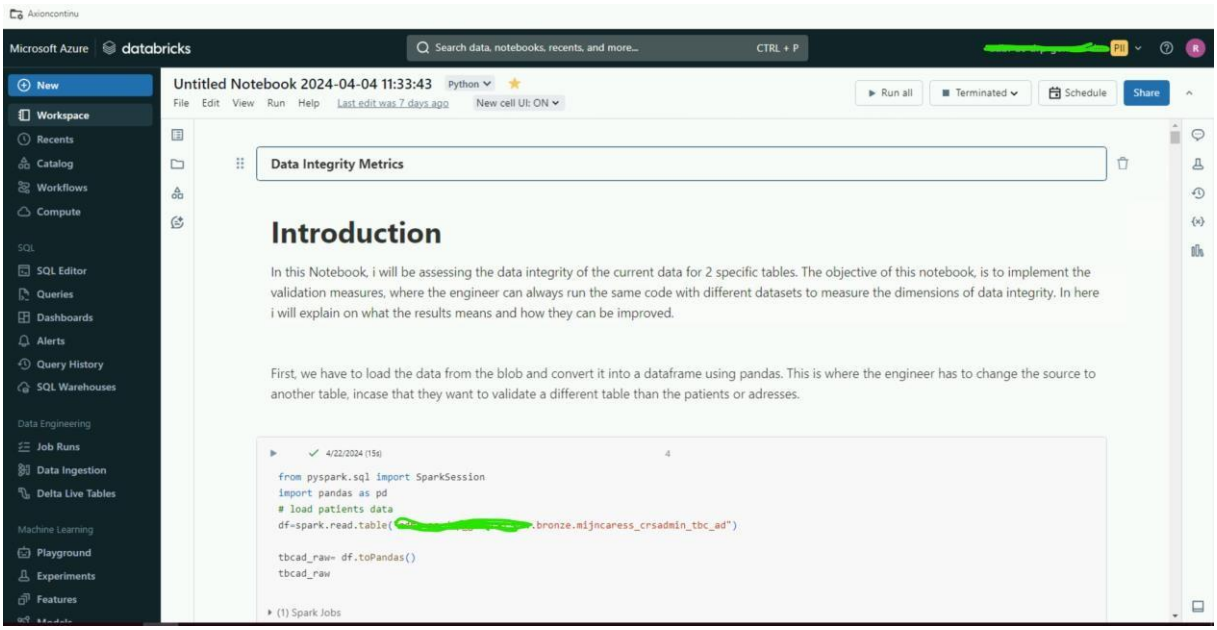


Figure 31 Screenshot of the notebook made for the metrics

8.9. Appendices I: Interview Linda2

Audiobestand

Linda 2nd interview 1.mp3

Transcriptie

Spreker 3

All right, it's recording right now. Yeah. So because, you know, like, for example, I don't know if you've seen one, one types of the datas that that I have.

Spreker 1 No.

Spreker 3

Like the patients, yeah.

Spreker 2

This is for my cars, yeah.

Spreker 3

From mine, one of the patients. Data, right? Like we have. ID's long now let me.

Spreker 4

Just put it there. Yeah. Oh my God.

Spreker 3

This is 2024. I'm so 2008. Just like the other day, I realised one of my neighbours son was born in 2007. Sorry 17 I I was like, how old are you now? Like 8. That's impossible. Hello. Spreker 1 Yeah. Spreker 5 Do it.

Spreker 1

Yeah, there.

Spreker 6

Oh, yeah, yeah, yeah.

Spreker 3

I did it.

Spreker 4

All right, so Mia. Mia.

Spreker
7

Egg habits. OK. Now. Yes. Thank you. As you can see like these, these are the types of, for example, this is just.

Spreker 3

The first table right.

Spreker 1

Yeah, yeah.

Spreker 6 Now.

Spreker 3

Like earlier from what we heard, yes, we're saying, but what I from what I've heard. About Jasper saying is that.

Spreker 1 Next.

Spreker 3

This data needs to be checked over with quality like for example in the names about the spaces, etcetera. But to Michael, this is not interesting.

Spreker 1

Yeah. Yeah.

Spreker 2

Yeah, Michael, once high over, you know like.

Spreker 3

What is the high?

Spreker 2

Over over, he wants to check for example data called. You can check already from the source. Like are the data types not changed for example, so that's what I check in the Azure data factory, I build pipelines and then I check like column X, the initial data type is int and if it changes to a Boolean sample.

Spreker 3

OK. Yeah.

1

Spreker
Yeah.

Spreker 3

Will you know something? Yeah.

Spreker 2

Then the pipeline will fill and then you'll see. Oh, they change something in the data type, or maybe they change more in the source and that's why the data is not reliable. That's important data, call it check and also like.

Spreker 6 I

see.

Spreker 2

The number of rows. For example, if you look at the source that's bronze, it's the source, and if the number of records has changed is if the number of records is the same in bronze as in silver R for example, those things. Also you can check. You can also check like some columns.

Spreker 3 OK.

Spreker 2

Didn't have no values. So like like. No, no, not the spaces, but some some. I think some columns have like a space in it, and then it's actually empty as well. So those things, yeah, those things you can put out too and some columns have like AT or F.

Spreker 3

No. Yeah. Thank you. Other question, Mark, apparently. Yeah, that is dumb. That's.

Spreker 2

And then you can make it 2 false actually, so that's also what I'm doing for for my cars in the big sofa. I'm I'm writing a script to detect if there's a team, make it too false and what more.

Spreker 5

True. False, yeah.

Spreker 1 OK.

6

Yeah.

Spreker 3

Spreker
See Saint George.

Spreker 2

U. M. Changing the column names because it's not nice to have the lower case and stuff the the last scrape you know, and so that needs to be removed. The location and the names are all in uppercase, so they're they're all needs to be.

Spreker 3

Yeah, only in uppercase.

Spreker 2

Pascal case, so both adults will change the keyboard to that and harborton will be better and that and all those things. But what? What did? Michael says that he's interested more in.

Spreker 3

How do you? Yeah.

Spreker 6 I

see.

Spreker 3

You know, like for example, if we look at consistency.

Spreker 6

Right. Yeah, in.

Spreker 3

Long names, for example. That's not interesting to him like.

Spreker 1 Yeah.

Spreker 3

Names or names, you know? Yeah, I was like, yeah, you could look at the uppercase. Lowercase. Yeah, but nothing much. No.

Spreker 2

It's not the most quality issue. But the most the quality is like this is the sort. This is the source and maybe we have another source. I don't know much about. If there are other source. If you want to check whether whether all the records are the same from this and this.

Spreker 4

I have no idea, yeah.

Spreker 3

Same as in there that this is also hard to check because I really I don't have any other sources right and like and I'm supposed to look at just the bronze layer for example of this, right? So it's really hard to compare the rows with something else, right? That's for one also like.

Spreker 1

Yeah. Yeah. Yeah, yeah.

Spreker 3

What was it again? Uh, yeah. Like, yeah. So if we looked at the, uh, yeah, uniqueness. For example, if we look at the uniqueness at.

Spreker 1 Yeah.

Spreker 3

Let's say the first column. This doesn't matter to him because yeah, he expects it to be already unique, you know.

Spreker 1 Yeah.

Spreker 2

You cannot expect you have to check out first expect with maybe, but did you try to check the uniqueness of for example the?

Spreker 3

What I mean? Unique. So why waste time on it? So like?

Spreker 2 1st.

Spreker 3

Yeah, I did it. It's all unique. It's all unique like.

Spreker 2

And and and it's. OK. Yeah.

Spreker 5

It's also.

Spreker 2

But is it a? Is it a primary key? Because if it's not the primary key, then.

Spreker 3

Yeah, this is the primary key from the tab of this table. Yeah. And this is also unique IDs, just per patient, right? Yeah.

Spreker 1

Yeah, yeah.

Spreker 2

So some based on the data. Object structure. You know you can based on the meta data of that you can see if it's unique or not already, yeah.

Spreker 3

Yeah, that's true. I see. Like for example, this is the Excel sheet that well, you know this excel sheet, but I changed my part, right? So for example this additionally quality controls like you see the IC a number. Nothing, because nothing interests him there, right? Like no name. Maybe the consistency for capitalizations. For now, I'm also, like, capitalization. What's not next? Yeah, then now it comes to the there was something else where we had a little check, like 11 proof tests.

Spreker 1

Yeah, yeah. Yeah.

Spreker 2

Yeah, but the version, for example birth serves number. That's it. Yeah, yeah.

Spreker 1 Yeah.

Spreker 2

Yeah, yeah. And if it's unique as well per person.

Spreker 3

Right. Yeah, but did that to him. It's. Yeah. I don't think Michael is unique. This. It's that way, you know, like, well, what's the purpose of measuring uniqueness? Here, that's my question.

Spreker 2

Yeah, because better service number it cannot take. Yeah, logically you cannot have the same perception as me, so, but that's that's also an assumption. Maybe there is something wrong in the system and they did. You don't.

Spreker 3

Yeah, but this is a given for. You don't know. Yeah, that's true. Yeah, that's.

Spreker 1 So.

Spreker 2

So you can assume that it must be, but I'm not of the assuming thing.

Spreker 3

You know, Sam, I'm not a fan of the assumptions.

Spreker 2

I want to check first but is this? Is this really so? Just check it by by the script. Great. And then, yeah, yeah, it is unique in this.

Spreker 6 Right.

Spreker 4

Right now see.

Spreker 3

You mentioned like, yeah, you you could do like a couple of normal analysis. Yeah, just very normal analysis, which also just doesn't include like the D types. It would show most of them.

Spreker 1 Yeah.

Spreker 3

As objects.

Spreker 1

Yeah, yeah. Spreker

3

Or date type. So how do you actually see if it's? If it's an integer or something, you.

Spreker 2

Know. Ohh you can I think if you go to the I I know somewhere you can check more about the data let me check.

Spreker 5

All I see is options.

Spreker 2

If you go to. Example here. Normally you you you see here.

Spreker 3

Now. Would see it. Nice.

Spreker 4

But you can't see it in.

Spreker 2

A script? Yeah. You can make a script for getting the data data each object.

Spreker 3 Right.

Spreker 2

All right. All right. You can make them a table or something based on the script you insert it in the data frame. Then you know from this this table has these columns and these columns. Yeah. Yeah, that's.

Spreker 1 Hey.

Spreker 4 Not

third. Spreker

3 And that's a

different way

to check.

Spreker 2

Yeah, to know what's the initial data type and stuff and how many records are in there.

Spreker 3

Well, for example, like. Yeah, yes, you could see these are actually unique, right? Like this the count, this is an actual unique right etcetera, right? But yeah, to Michael, this is all.

Spreker 1

Yeah. Yeah, yeah.

Spreker 3

Eda General Data exploration, right, that's like, yeah. But how would you consider measuring uniqueness in this?

Spreker 2

Yeah, yeah, yeah, yeah. Scenario uniqueness per column or per.

Spreker 6 Like.

Spreker 3

Row. Yeah, but what is the purpose per column? That's it.

Spreker 2

Yeah. For example, primary keys and foreign keys. Those are the things that should be unique. Per row, and if they're not unique, then yeah, you've got. Problems.

Spreker 3

And per column.

Spreker 2

And per column. Yeah, sometimes per column is not unique. For example, I remember for Athos, we have a column loan coaster and there are sometimes you have one, one made of Archer five times because made of Archer you have also a column of periodic period 123.

Spreker 1 MHM.

Spreker 2

So it's normal that you see multiple times, so it per column. It doesn't have to be unique but say per row it has to be because then you based on the row and all the other columns you can create.

Spreker 3

Both, yeah. It does because of the time. Yeah. Yeah, I know.

Spreker 2

Example, another unique hash hash value or something.

Spreker 3

You called the ID or something cash level. Yeah, yeah, yeah, I see, I see, like.

Spreker 2

And do you also check on duplicates for example?

Spreker 3

Yeah, but that's totally part of the uniqueness, right?

Spreker 2

Yeah, that's part of the uniqueness per row. Yeah, if they're the double rows, then they they're not the stages. When you make a dimensional model, you like, push it. Something goes wrong because it's not unique.

Spreker 3

No, that's true.

Spreker 6 Yeah.

Spreker 2

The uniqueness is important, and if it's if it's all the data you want and the data quality is like. Spreker 4 Yeah.

Spreker 2

Yeah. What's what is in each column? Because some I see for my cars, some columns have zero values, you know. And now in the my script that I'm writing, I'm like analysing each. Spreker 1 MHM.

Spreker 2

It's stable and if it's filled with nothing like literally nothing, we we philtre that column out. There's you don't. Why would you use like a table with 10 columns and like 6 columns 0 you don't need. Yeah.

Spreker 6

I see, yeah.

Spreker 3

Empty. Yeah. Yeah, that's true. Yeah. Yeah, that's very true. Like see. Like OK, What I'm trying to get at is like for example, consistency could be measured for every single column in this table, right? Like. Yeah, for for telephones, you could always measure the consistency of numbers. That starts with +31 or 0031, etcetera, etcetera. Now you need.

Spreker 1

Yeah, yeah, exactly. Yeah.

Spreker 2

And also like with how many numbers should be in? Yeah then.

Spreker 3

Yeah, length per. Yeah, etcetera. But this all falls under consistency right now. What falls under accuracy?

Spreker 1 Yeah.

Spreker 2

Accuracy is. Yeah, I think I have it in mind, but how do you know that the data is? Accurate. Let me think accuracy. I thought a crisis up to date. That the data is up to date and we in bronze we have our input time and that should be today.

Spreker 3

I see, yeah.

Spreker 2

And then silver, we remove that column because we don't need, we need 10 bronze to see how accurate the source is per.

Spreker 4 Yeah.

Spreker 3

Data. Yeah. Per day, yeah, but there is no way of measuring accuracy. In here, per column or per Roth?

Spreker 2

There's a bronzer or what is this?

Spreker 3

Yeah, yeah, this is bones. Is literally there from words, but I picked specific tables, yeah.

Spreker 2

You only tricks called. Yeah, well, you can look at the input input time. Yeah, input time is not today. Then it's not accurate. Yeah.

Spreker 3

Then it's it's not access. Yeah, let me. Alright, that's a. That's a good one. And do you consider accuracy as part like for example the error that we found the other day like where date of birth was 10 years after the death or 20 or 50 is that accuracy or is that errors?

Spreker 2

Oh, yeah, yeah, yeah. That's. Reliability, reliability, but toolbar? It's the data. It's not reliable. Yeah, not reliable.

Spreker 3

It's not reliable because it's just part of accuracy, because accuracy also tells you how true the data is.

Spreker 2

It's part. Up to the data and how we send the data, yeah.

Spreker 3

And how do you? So it is part of accuracy.

Spreker

Doesn't seem.

Spreker 2

But but I'm thinking accuracy, reliability so many.

Spreker 3

Yeah, reliability definitely.

Spreker 2

Things. But how? By rights, intrigue sites. Integrity.

Spreker 3

Integrity is accuracy, completeness and.

Spreker 1

Yeah, there.

Spreker 2

So it all belongs to integrity. Actually those things that the day that what you said about that issue.

Spreker 3

It's it's data integrity. All right, all right, I see.

Spreker 1

Yeah, yeah.

Spreker 3

Because yeah, basically I am looking at data integrity, which fills completeness, accuracy, uniqueness and.

Spreker 2

Yeah. Then you test like a birthday. It should always be.

Spreker 7 Yes.

Spreker 2

Before that it works and you can see which cases it deviates and.

Spreker 3

Yeah, literally. Yeah. Hmm. OK. That's fair. Yeah, that's fair. And how how would you also measure completeness here like non values or null does not mean that it's not complete.

Spreker 2

No, because some columns are allowed to have no values, so complete this actually we have all the data.

Spreker 3

How do you actually measure? What do you mean by all the data?

Spreker 2

All of it, like all those, if you know, like in the source, there are like eight 1.8 million records. If you see that in the metadata, then based on the metadata you should check OK is the number of votes that I have the same as the numbers of votes coming from the method out of the source? Yeah.

Spreker 3

Yeah. Then it must be, yeah. Matter. That is that the only way to. Measure the completeness in here.

Spreker 2

Is it complete? Does it has all the columns from the source? For example complete? So all the rows. Columns. And also. Yeah, it's all the data in there because you can have all the rows, but it's. If the data in the rows is not all filled like in the source, then it's also not complete. Spreker 3

I see, yeah. So completeness can only be measured if it is compared with something.

Spreker 2

Else yeah, you need it. Compare it with the source actually.

Spreker 3

OK, that's a very good one. Honestly. All right. And what about? Yeah, we talked about consistency, accuracy, uniqueness, right. Yeah, yeah, yeah, the the only thing. The only way we could measure uniqueness is.

Spreker 1 Next.

Spreker 3

Primary keys and IDs in per row. If every row is unique, right?

Spreker 2

Yeah, yeah, exactly. And if every row should is unique, you should not have duplicates.

Spreker 3

OK. Yeah. And these are stuff that because, OK, these will also help in developing the ATL, developing the protocol for the ATL processes. So is this stuff that you actually keep in mind when you do the ETL?

Spreker 1

Yeah, yeah, yeah.

Spreker 2

Processes. Yeah. Look at. OK, I look. Always at the metadata. Like what do I expect to? Yet in terms of data types, in terms of number of records and stuff, and in terms of number of columns per for each object for each table I look at that and based on that I define rules already in the source like OK every day. New source data is loaded, so these source data in a pipeline that I've built, I check on the data types I check on the schema. Is it still the same if things sometimes things change. I added many times also at the Rabobank that the people in the from the source will provide the source data that they just suddenly change something without communicating and everything and the rest of the posters went wrong. Error, error and then you look back at the source and hey, did you change that?

Spreker 6

The salad. English.

Spreker 2

Yeah, we actually migrated this part of the column. So that's yeah. So in the beginning you should already check the how do you say the completeness and that it and the integrity isn't the data changed. As compared to what you defined of the source and define it with the schema and with cheque on the data types.

Spreker 1 Yeah.

Spreker 3

Alright, see I did some as I mentioned earlier, I did some consistency tests based on all the columns right or the chosen columns and these are this what micro and I decided to maybe make it this way. So basically the ETL advice I would love for you to validate my ETL.

Spreker 1

Yeah. Mm-hmm.

Spreker 3

Devices like the difference between output device and detail device. The output device is given to them right, like share and continue you guys this is advice for you. You guys do with whatever you.

Spreker 1

Yeah, yeah. Yeah.

Spreker 3

That's right, but ETL advice. It would be very nice if we. Could validate it.

Spreker 2

Validate the existing ID's, ensure that they need the specific length and format before loading to the database.

Spreker 6

Consistency.

Spreker 2

So check on the the length, yes.

Spreker 3

As lens was.

Spreker 2

That the length of the source and the source the length is already defined, yeah.

Spreker 3

Fine, yes. But yeah. Is it before you load it? Like actually transferred it into bronze? Let's say yeah, while you're loading it, you check it again, you know.

Spreker 1 Yeah.

Spreker 2

While while I just loaded from it's a SQL database, I load it in.

1

MHM.

Spreker 2

In the BLOB storage, OK. And then I load it here in the bronze thing and then I'm actually building some cheques.

Spreker 6

You that is. I see.

Spreker 3

All right, so you're doing it in the silver cheques?

Spreker 2

Yeah, they are. Not before the the balance is like raw raw data. They don't change anything and.

Spreker 3

Not. Alright. Alright, alright. I see. Yeah, that's true. That's true. Fair enough. Yeah. Then I'll change it into silver.

Spreker 1 Yeah.

Spreker 2

And the silver eye. You're going to check like the length. The length must be this or is the lens. And there you can also see OK, some some doesn't have, some have more more lengths or some have. Yeah, no length, or. Yeah. So in silver eye do these things.

Spreker 3

Database, right? In case of ID exceeding the limit flag, the row dropped from the main data set and validate it makes sense.

Spreker 1 Yeah.

Spreker 3

Alright, long names or long names and any name any table in column that contain names.

1

Spreker

Yeah.

Spreker 3

Yeah, PP ETL advice. Split full names into separate columns first.

Spreker 2

Last, etc. First name column, last name, Cologne. Yeah, but yeah, that's something small, actually, yeah.

Spreker 3

Yeah. Yeah. Instead of making it into one long. Yeah, but what else? I I don't know how to validate consistency for names other than uppercase. Yeah, check for uppercase. Spreker 1

The. Yeah.

Spreker Oops.

Spreker 3

Uppercase, but nothing else, right?

Spreker 2

Starts with. It starts with big letter indeed the name and the and the other name too deed.

Spreker 3

For an R validating existing IDs, yeah, it's the same thing as the one above because they're both. Unique numbers, right?

Spreker 1 Yeah.

Spreker 3

To the fossils, validate the data on entry log. These error. Uh yeah these errors and find the cause of it and treat the cause. Clean the data in case of exceeding the limit. This is based on an error that I found. Which is? This was part, yeah. Apparently Val was his last name. I don't know why he said a dash after it. And there, I don't know if this is a part of fossil. It was very weird.

Spreker 2 I'm

rival. 1 OK.

Spreker 2

I think from the wall sounds more, but yeah, no, yeah.

Spreker 3

Yeah, I thought about it, but none of Val is a last name, but it was still in the in the same column as.

Spreker 1 Yeah.

Spreker 3

This full source.

Spreker 2

Ohh that that's weird, yeah.

Spreker 3

Exactly like validate on entry lock. These errors find the causes may be treated.

Spreker 1

Yeah, yeah.

Spreker 3 Yeah.

Spreker 2

There are many things to improve in the data itself, because now it's silver I I started with it and I I started with. The first thing is like changing the column names so they're all uppercase. Every letter is uppercase and you have the last paper. So I want to make it normal names.

Spreker 6

Yeah. So because. Yeah.

Spreker 1

Hmm.

2

More.

Spreker 3
Spreker

Standardisation.

Spreker 2

And yeah, and all the Boolean types that we see like they have. I want to make them two false as yellow and one as well. So that's what I did.

Spreker 6

Alright, yeah.

Spreker 3

Same as here for example the slat. I've noticed that it's just M. And F right.

Spreker 2

Column as it did.

Spreker 3

Yeah. Yeah. I was like, in case of 1/3 value you observe, yeah.

Spreker 2

Ohh. Oh man, male feet. Ohh, we'll we'll change it. I remember that I talked to the hospital. It will be M capital letter and they.

Spreker 3

That's the e-mail. Yeah, alright. Validate the row. Consult the business side application side on the data entry to ensure only valid values in the column. This is in case that. Spreker 1 Yeah.

Spreker 3

The third value was observed.

Spreker 2

Yeah, they should, actually. It starts already where the source data comes in. So they say they should have input the controls that you only can or maybe a drop down list or that you cannot. Yeah.

Spreker 3

You standardise.

6

Yeah.

Spreker 3

Dumb list is good.

Spreker 2

Because if people can just type in things, there goes the press or yeah.

Spreker 3

Yeah. Literally, as here. Also, as I said, use standard values to ensure the consistency for their genders. Either continue with using F&M or. Switch to use male and female like seriously. Yeah alright.

Spreker 8

Yeah, yeah.

Spreker 3

ETL advice. Validate the dates within a reasonable range. And follow the. Expected format for example DOB or the IS before DoD in case of data exceeding the limit and errors to be observed blah blah blah.

Spreker 1 Yeah.

Spreker 3

Sounds legit.

Spreker 2

Yeah, right, indeed.

Spreker 3

Alright, board plats advice is to ensure consistency and quick access remote. The futuristic by normalising the data before loading into the database. This is, yeah, like accents or names, right? Like for example, because you've seen in report places you've seen settle them both seen them both I've seen.

Spreker 2

Ohh yeah yeah.

3

Yes.

Spreker 2

You see, it really set off of those dimples, so it's not standard.

Spreker 3

Yes, yes, that's not standard. No, I've seen Utrecht, I've seen. Ooh.

Spreker 1

Lift like ohh wow. Yeah, yeah.

Spreker 8

OK. Mike, OK.

Spreker 2

From where you from, there's even the small village, yeah.

Spreker 3

I know that there's Eric, but the the only thing I think is is everybody think they literally have their thinking sound, I don't know.

Spreker 1

Yeah, yeah, yeah. Oh wow.

Spreker 2

Just really straight trash. Garbage in garbage.

Spreker 3

I don't know. Have you ever been too threatened? Screamed in the station or something?

Spreker 2

Not yet. I worked at it, but I've never been to you.

Spreker 3

But you know what I mean? Like, the first time I heard about it.

Spreker 8

Yeah, I know.

Spreker 3

Yeah. So this explains, yeah.

2

And they also share this with me because it's very done. I can also use it with my thing. I want to do.

Spreker 8

So my stuff is being used. Yeah, yeah. That's.

Spreker 3

I guess I'll share it. I also made something else. I don't know if this is going to be useful for you. It's called, yeah, data integrity advice. Protocol. Which in it includes like for example the defaults. Let me make it better the definitions of what is complete, what is criterias?

Spreker 1

Yeah, yeah, yeah.

Spreker 2

So somehow I know it in my head, but it's good to look at what do. They really.

Spreker 3

I mean, yeah, like completeness complete patient data includes all necessary information, tactically assess, plan, deliver, whatever, etcetera, etcetera. A lot of information, roles, responsibilities, data collections, dation what what could be very interesting for you is the last chapter.

Spreker 1

Yeah, yeah. Yeah.

Spreker 3

Maybe this the business rules and metrics like for example yeah. Check the required fields, specify which columns are important for decision making and others complete this, etcetera, blah blah blah. And then here's some metrics that they also included what you.

Spreker 1

Yeah. Yeah.

Spreker 3

Could do to here. Yeah, and same goes for the rest accuracy, uniqueness.

5

Inconsistency.

Spreker 2

Interesting. Yeah, consistency. And also like the thing that you saw from you and stuff we must discuss with the business, I mean, because I can assume that it's you or or something else or some strange places. And so those things.

Spreker 4

Yeah, basically.

Spreker 8

All right.

Spreker 2

To be discussed with the business.

Spreker 3

Definitely this is.

Spreker 2

And maybe for them that they know they have to change it from their side because yeah, it spreads a lot of work on our side if they just.

Spreker 6

Yeah, and.

Spreker 2

The the **** together. Just yeah.

Spreker 3

And literally and literally, she's together for. Yeah. Yeah, that. See. So that that is that right. That was for the first table. Now I'm I'm kind of struggling with the second table which is the

addresses. Table, right? Yeah. Like OK, consistency on Strat. Now how am I to measure this so that names are supposed to contain dots? Like professor or whatever. Yes.

Spreker 2

From the yeah.

Spreker 3

It can contain numbers etcetera, so I don't.

2

Yeah, but start Nam, it's only it's mostly only only layoff on the earth. So I think it's numeric for sure. Straight name street name. Yeah, because it can be layered. That in there. Yeah.

Spreker 3

Right, but how do you measure consistency here? I.

Spreker 2

Don't know. Well, it ends with start for sure.

Spreker 8

Yeah, if you don't have this problem.

Spreker 3

Yeah, that now that's a good question. If they actually all end with strands.

Spreker 8

It's not as though.

Spreker 1

Yeah, yeah.

Spreker 2 Yeah.

Spreker 3

House number I don't know. Not too much. Or like maybe check for weird values of beginning with 0.

Spreker 2

Spreker

But because they have house number, but do they also have as a separate column?

Spreker 3

Yeah, they do. Yeah. Here. House House #2, right? Yeah, yeah.

Spreker 2

Oh, yeah, yeah. Two full. So ours number should only be numeric values and two, four you can.

6

Yeah.

Spreker 2

House number should be int I think.

Spreker 3

It should be an int, but house number it yeah int.

Spreker 6 OK.

Spreker 2

Ohh some house number 70.

Spreker 4 Or

4.

Spreker 2

Yeah, the zero. Yeah. Why would you use it? Yeah, yeah.

Spreker 3

So I could measure to 0. Her consistency.

Spreker 2

Somehow like 01? Yeah. Oh yeah, 01.

Spreker 3

But is it possible that some houses?

Spreker 2

Spreker

If you make the ends then it will be one. That's the thing.

Spreker 3

Could have 0. Yeah. Do houses actually exist? Like 02 house #0? Not house #2.

Spreker 2

I know two or two, or zero or one, but house number I've never seen. But.

Spreker 5

I've never seen.

Spreker 2

Probably hear this.

Spreker 3

Two things with the number I know how to.

Spreker 5

Put this with ABC.

Spreker 2

When I lived in Rotterdam, I had 2415. No, say or something very weird. Someone just said a big house and made small studios in it, you know.

Spreker 5

Nice. Yeah. Yeah. That's why we have like ABC.

Spreker 8 Yeah.

Spreker 3

Yeah, it's it's just weird. But yeah, I get it. Awesome. Yeah.

Spreker 2

But what do we do? Because it's house #0 in front. You can make a int. But technically it's done not how it's house numbers divine defined, so I'm adopting instead of int. It should be actually into house.

Spreker 3

Yeah. That's true. Yeah. I could also check if if any of the rows contains any letters in it.

Spreker 2

Yeah, yeah, indeed.

Spreker 4 Uh.

Spreker 3

Yeah, contaminated with letters. All right. Uh postal code. Yeah, I yeah. Consistency is you could check.

Spreker 2

Should have, I think 6. It should have six, yeah.

Spreker 3

Or five. 6-7 I think 7 length.

Spreker 2 Five

day.

Spreker 3

Yeah, if you're getting the space between.

Spreker 2

Or sometimes there's space, sometimes there's.

Spreker 3

This is not now. See, I've done the analysis. You'll see it. It's really weird. Honestly. Hold up. Ah, chaser.

Spreker 2

So it's also not consistent at the post.

Spreker 5

Code ohh it's crazy.

Spreker 1

Oh my God.

Spreker 5

I swear to God when I swim.

Spreker 2

They don't have input controls at all. People this can, and that's why you have this inconsistencies.

Spreker 7 Sadly.

Spreker 4

Yeah, here. Yeah.

Spreker 2

So we also need to make that consistent, like if the post code has a space in between or not.

Spreker 3

Yeah, these are all the post codes we have. These are all the different lengths available. You either have one or eight. One is a space as you see, not even empty.

Spreker 1

Ohh yeah yeah yeah.

Spreker 3

You have 4 which is none. Yeah, you have this one, which is like just a number which for example in. Yeah, some countries they don't have letters after of code. So makes sense right.

Spreker 1 Yeah.

Spreker 2

Yeah, it's true. Yeah. But this is from the only Netherlands people.

Spreker 3

I don't know mine, Chris. I don't think it's just Netherlands. I think that's also part Frank.

Spreker 2

OK. Ohh yeah, that's the thing. Yeah. Yeah, yeah.

Spreker 3

Far somewhere. So I bet this is French 100% no joke.

Spreker 5

These two or these three are Dutch, but you could see the inconsistency space space.

Spreker 2

Yeah, there with the. Yeah. So what's the choice? What do we do? Do we? I think without faith.

Spreker 6

And a different letter.

Spreker 3

You know what this view is?

Spreker 8

Ohh, you actually it's also yes it's it's.

Spreker 7 Yeah.

Spreker 2

Check the post.

Spreker 3

It's like it's like hilarious.

Spreker 5

There is a uh, what is this?

Spreker 8 Well.

Spreker 3

Oh yeah, these. Are all the different types of? Uh. Traditional post code boom plots. Spreker

2

Ohh yeah. Spreker

3

But this is for example not interesting to Michael to see this. Maybe this is interesting. You know what I mean?

Spreker 2

Yeah, yeah, the post code. We keep them consistent.

Spreker 3

Yeah, von von plats. I can maybe only check capitalization. Yeah, completeness. I cannot check unless I do it with another table at a different source. Accuracy can only be measured with time input time and.

Spreker 1

Yeah. And Texas? Yeah.

Spreker 3

Yeah, errors like just errors like the date of birth stuff, data integrity.

Spreker 1

Now.

Spreker 3

And what was the last one? Completeness. Completeness also cannot be measured unless we have a different data source or that we actually complete the.

Spreker 2

Completely set. But you can completeness. For later. So duplicating. Somehow also, because sometimes you think you have 1.8 million records and when you check on duplicates it's actually 1.2. So 600,000 more and in the soil, but you need also the source to see if yeah so.

Spreker 6 Oh.

Spreker 3

Oh. Yeah, compare.

Spreker 1 Yeah.

Spreker 3

That's fair. I mean, uniqueness. I also cannot read. Yeah, it's only duplicated records, but you also need a different source to compare.

Spreker 2 Yeah.

Spreker 3

So the my best friend in this whole project is consistency I.

Spreker 1 No,

no.

Spreker 3

I don't like it.

Spreker 2

Because there's so many inconsistencies.

Spreker 3

No, it's no, it's just I want. To look at other. Dimensions as well.

Spreker 1

Yeah, yeah.

Spreker 3

No, because this means if consistency is the only dimension, I could look at, then that means I'll need to do the same for another table instead of looking at another dimension on the same table.

Spreker 1

Yeah, yeah.

Spreker 6 Yeah.

Spreker 3

So the work is just changing, that's fine, but I need to discuss this with Michael, then the the the rule seems OK. So I'll just share this document with you and the screenshot of the notebook. Yeah, definitely.

Spreker 1

Yeah. Yeah, that would be.

Spreker 2

Nice. Yeah. And the thing you. Yeah, no. Then I can use that also for.

Spreker My.

Spreker 8

I was also. Spreker 2

Always happy when they use my word, yeah. Then you do something useful, you know.

Spreker 3

Yeah, that's fair. That's fair. Yeah, it reminds me of my first internship that I've done, made a whole start for the dashboard and et cetera. And now that the whole company uses my dashboard.

Spreker 1

Yeah. Yeah. Ohh wow. Awesome man. 8

Spreker
I know, I know all.

Spreker 3

Right. Yeah. Then I guess that's it. Just have to validate the rest.

Spreker 1 Sure.

Spreker 2

And share it with me and later we can maybe also if you find something else you can always discuss, Sir.

Spreker 3

I'm going to discuss this.

Spreker

With Michael.

Spreker 3

Right. Yeah. And yeah, sounds good. Sounds good. Then I'll just have to ask, maybe ask her to give me also another table to look at. Yeah. But uh, yeah, yeah, we can.

Spreker 2

The 48. Tables.

Spreker 3 And.

Spreker 5

Our 48 kHz Mihos.

Spreker 7 Yeah.

Spreker 3

But there's nothing to do, for example, with the. Yeah, still happening, right? Like the dots. The first, the ears. Whatever they do, no.

Spreker 2

The thoughts, or do they have thoughts in the start line their dots?

3

Yeah. Yeah, like. Instead of professor.

Spreker 2

Spreker
That's really weird.

Spreker 3

Yeah. And sometimes they could actually also write it as professor.

Spreker 2

And if you Google this track names.

Spreker 3

I think it would be with Prof Dots.

Spreker 2

You see them with profiles also.

Spreker 7

I think so, yeah.

Spreker 1

That's all.

Spreker 3

I need to validate it double.

Spreker 2

Why didn't you just use a dropdownlist with all the street names in the Netherlands updated? Then the consistency you know.

Spreker 3

I don't know.

Spreker 2

But yeah, those are small things. I think that's happening and stuff.

Spreker 6 Yeah.

Spreker 3

But we should just leave it as it is. Then guess.

Spreker 2

Yeah, I think Bashan is very important to. I don't know if there's major records, but Bayesian is already unique. It's with that you can get everything out. Number.

Spreker 1

There. Yeah.

Spreker 6

Check range.

Spreker 2

And the post code thing is also weird. Sometimes with space, sometimes not. Not really consistent.

Spreker 3

Sometimes an extra letter.

Spreker 2

Yeah, it's important. Actually, the combination of number of the House and post code, if that's not correct normally. When you fill things in in the post code and. Number not correct, you get an. Error.

Spreker 3

That's fair. The in the application you mean? Yeah.

Spreker 2

Yeah, in the application. But then there's I.

Spreker 1

Don't know, I don't.

Spreker 3

Know I don't, plus I. Mean. I don't think it would be wise for us to check every row. Isn't there a library that does that?

Spreker 2

No. Oh, hell no, no.

Spreker 4

That's not bad job.

Spreker 3

Swear to God. Isn't there a library that does that?

Spreker 2

It should actually. They should make input controls in the source so it cheques.

Spreker 1 Yeah.

Spreker 3

Yeah, that's true and that's why I have. The application advice as well, yeah.

Spreker 2

Yeah, that's smart.

Spreker 3

Because I really at first I found it really hard to differentiate between ETL processes and application advice, right? Because.

Spreker 1

Yeah, yeah.

Spreker 8

Both to me is part of the.

Spreker 5

Advice you know.

Spreker 2

Efficiency. No. But when you see many inconsistency and it starts. With the source. No, the source gives us the data with inconsistency and we can do ATL. Yeah, we can change it all, but is it efficient? No.

Spreker 3

Should yeah, no, no. Yeah, that takes a lot of calculating powers.

Spreker 1 Yeah.

Spreker 5

And a lot of money, obviously.

Spreker 3

But then I guess, yeah, I guess that was it for her. Thank you for your time.

Spreker

Spreker 2

Welcome and thank you for your input.

Spreker

8.10. Appendices J: Interview Jasper

Audiobestand

Jasper interview.mp3

Transcriptie

Remy: All right, yeah.

Jasper: So I hope I can show you something that huck made..

Jasper: Uh. I'm not sure because this is what he did. So here this is goals, dimensions and facts. So this is our goal.

Remy: Right.

Remy: Yeah, but he separated them based on what ? you know?

Remy: Mean like because if I'm getting correctly so that you get all the data into bronze very well and it's just a mixture of everything, right? Yeah. So how do you separate the fact from the dimension? How do you know that this is actually the actual table based modes?

Jasper: Yeah, mostly because.Clear and controller function and incident status dimensions. So they already showed us like this is. What we need. So they give you a small guidelines to follow basically, yeah.

Remy: Yeah, this is the current situation like this is what they have right now and we have.

Remy: that's cool.

Jasper: We made it.

Remy: Yeah, that's yeah, I thought that we had to, you know, map everything based on something.

Jasper: But then in our way and just check like, yeah. Yeah, exactly. That's why we because really. Yeah, we just got a lot of information from actually continue and we just have to, yeah, read everything and make our solution.

Remy: And for now, there's nothing done to ensure any of the quality stuff, right.

Jasper: ahm not really like there are minimum stuff. Very bare minimum.
Remy: so accuracy is also important.

Jasper: Yeah, yeah.

Jasper: I'm guessing accuracy is like the reason that they are coming to us because we told them we we can do quality checks on the data.

Jasper: All the data that we show will be correct. I see. Yeah. Yeah. That's the challenge as well because. Yeah.

Remy: Because I was also reading about the accuracy earlier, I mean there are methods to measure the accuracy to. Like one of the methods that you could quarantine the data that you get right where you could load in the good data and the bad data which it could flag. You could load it in A different. Database and then you ask about the different database and then you add it into the main one. You know, yeah.

Remy: OK. OK. So yeah, we said consistency, accuracy, completeness. Have you noticed anything within the data that is actually not complete at all?

Jasper: People and I've set up this small excel. But then so it's hard to figure out exactly.

Remy: This is why. I would need like the data so I could. And analytics on it, you know, distributions.

Jasper: Like sometimes there are like not null columns but they are empty like you cannot fill in not null so. It just make an empty version because.

Jasper: No, doesn't work, but the whole idea of nodes like that if they need to have a value so. There are some interesting. But yeah, I'm still working on. Yeah, understanding all the data, what doesn't mean? What do they expect? Why are there some columns not know why.

Remy: And now we. Yeah, the format you just mentioned there is a desired format and I'm guessing if it is already done by someone, right? Yeah. The desired format that you want it in, right? Yeah, you made it right.

Jasper: This is what I meant.

Remy: So there is, yeah, I could put it under should as validity because it's already there.

Jasper: Yeah, yeah. But to actually do some all of the sources, because this is only caressed, but we have more.And just have to research it.

Remy: And I mean, but also for the internship scope is just like, you know. Yeah, OK, what about timeline.

Jasper: Michael's working on that, yeah.

Remy: What is he doing for?

Jasper: It any ideas on the timeline now? Yeah, yeah. He's just because the scope is different. And a few months ago.

Remy: That. Yeah.

Jasper: This is making a new timeline, but when what is finished what what we have to do when yeah.

Remy: I see I.See, but I meant by the timeline of the data itself, but because, yeah, yeah. Because usually the timeline is like.Well.What you know we'll. We'll time stamps. When does it come? When does it leave? Compared to the consistency of the other sources.
But since we only do one source could.

Jasper: Yeah, it's.

Remy: Uniqueness. Yeah, I need the data to test these, you know, because you cannot know what is what is not.

Jasper: But I hope Michael has permission for you to get in the server acting on the new and see the data for. Yourself.

Remy: yeah, I Need some time to find out the tables and understand them.And it will take some time to you. Know. Yeah. All right.

Jasper: Then you can. Confirm what I find, or maybe add some suggestions like maybe this is better or.

Remy: Yeah.