Lotta Lassila

Developing a Method Artifact for Warranty Data Analytics

Enabling New and Reliable Insights Through Integrated Data Processing and Dashboard Development

School of Technology and Innovations
Master's Thesis in Information Systems
Master's Programme in Computing Sciences

Vaasa 2026

UNIVERSITY OF VAASA
School of Technology and Innovations
Author: Lotta Lassila
Title of the thesis: Developing a Method Artifact for Warranty Data Analytics: Enabling New and Reliable Insights Through Integrated Data Processing and Dashboard Development
Degree: Master of Science in Economics and Business Administration
Degree Programme: Information Systems
Supervisor: Timo Mantere
Year: 2026 **Pages:** 81

ABSTRACT:

Granting warranties can represent a significant cost for companies, making accurate warranty data analysis essential for effective decision-making. The company involved in this research has acknowledged this strategic importance, and its warranty management aims to further develop its capability to utilize the full potential of available data. From this perspective, a new need has emerged to combine warranty data with climate data, which has not previously been included in the analysis, while also placing strong emphasis on reliable data quality. Increased investment in data utilization has also led to the adoption of modern data architectures, where computationally intensive data processing is separated from visualization in data analysis.

The objective of this thesis is to develop a repeatable method artifact to create an integrated view of internal warranty data and external climate data to support warranty management's decision-making. The practical components of the artifact combine Databricks and Power BI, where Databricks is used for extract, load, and transform (ELT) process and data quality management, while Power BI is employed for visualization in the form of an interactive dashboard.

The methodological approach for this study is design science research (DSR) which focuses on solving organizational problems by designing and assessing innovative IT artifacts. Within DSR, a method artifact is a structured process that provides instructions for solving a specific problem. The method is developed by combining information from prior research with insights obtained from domain experts within the organization and is designed through rapid iterations. By combining prior research, domain expert insights, and iterative design, the method ensures alignment with the rigor, relevance, and design cycles central to DSR.

The method is demonstrated in a real organizational context by comparing the proposed proof-of-concept method and the resulting dashboard against the defined requirements. The results of the demonstration indicate that the method succeeds in addressing the identified problem. Furthermore, the findings show that the method enhances decision-making by providing the desired insights in the dashboard with reliable and trustworthy data quality. However, the generalizability of the method may be limited by the single-organization context, and the proof-of-concept implementation might constrain certain aspects, pointing to opportunities for future research.

KEYWORDS: business intelligence, data processing, data storage, visualisation, warranty

VAASAN YLIOPISTO
Tekniikan ja innovaatiojohtamisen akateeminen yksikkö
Tekijä: Lotta Lassila
Tutkielman nimi: Developing a Method Artifact for Warranty Data Analytics: Enabling New and Reliable Insights Through Integrated Data Processing and Dashboard Development
Tutkinto: Kauppatieteiden maisteri
Oppiaine: Tietojärjestelmätiede
Työn ohjaaja: Timo Mantere
Valmistumisvuosi: 2026 **Sivumäärä:** 81

TIIVISTELMÄ:

Takuiden myöntäminen voi muodostaa merkittävän kustannuserän yrityksille, minkä vuoksi tarkka takuudatan analysointi on olennaista tehokkaan päätöksenteon kannalta. Tässä tutkimuksessa mukana oleva yritys on tunnistanut tämän strategisen merkityksen, ja sen takuutiimin johto pyrkii kehittämään kyvykkyyttään hyödyntää saatavilla olevan datan koko potentiaalia. Tästä näkökulmasta on syntynyt uusi tarve yhdistää takuudataa ilmastodataan, jota ei aiemmin ole sisällytetty analyysiin, samalla korostaen luotettavan datan laadun merkitystä. Datan hyödyntämiseen kohdistuneet kasvaneet investoinnit ovat myös johtaneet modernien data-arkkitehtuurien käyttöönottoon, joissa laskennallisesti raskas datan käsittely on erotettu visualisoinnista data-analyysissä.

Tämän tutkielman tavoitteena on kehittää toistettava menetelmäartefakti, jonka avulla voidaan muodostaa integroitu näkymä sisäisestä takuudatasta ja ulkoisesta ilmastodatasta takuuhallinnan johdon päätöksenteon tueksi. Artefaktin käytännön toteutus yhdistää Databricksin ja Power BI:n, jossa Databricksia hyödynnetään poiminta-, lataus- ja muunnosprosesseissa (ELT) sekä datan laadun hallinnassa, kun taas Power BI:tä käytetään visualisointiin interaktiivisen dashboardin muodossa.

Tutkimuksen metodologisena lähestymistapana käytetään suunnittelutieteellistä tutkimusta (DSR), joka keskittyy organisatoristen ongelmien ratkaisemiseen suunnittelemalla ja arvioimalla innovatiivisia IT-artefakteja. DSR:n mukaisesti menetelmäartefakti on jäsennelty prosessi, joka tarjoaa ohjeet tietyn ongelman ratkaisemiseen. Menetelmä kehitetään yhdistämällä aiemmasta tutkimuksesta saatu tieto organisaation asiantuntijoilta saatuihin näkemyksiin, ja sitä kehitetään nopeiden iterointien avulla. Yhdistämällä aiemman tutkimuksen, asiantuntijatiedot ja iteratiivisen suunnittelun varmistetaan menetelmän yhdenmukaisuus DSR:n keskeisten rigor-, relevance- ja design-syklien kanssa.

Menetelmää demonstroidaan todellisessa organisaatiokontekstissa vertaamalla ehdotettua proof-of-concept -menetelmää ja siitä johdettua dashboardia asetettuihin vaatimuksiin. Demonstraation tulokset osoittavat, että menetelmä onnistuu ratkaisemaan tunnistetun ongelman. Lisäksi tulokset osoittavat, että menetelmä parantaa päätöksentekoa tarjoamalla dashboardissa halutut näkemykset luotettavan ja korkealaatuisen datan pohjalta. Menetelmän yleistettävyys voi kuitenkin olla rajallista yhden organisaation kontekstin vuoksi, ja proof-of-concept -toteutus voi rajoittaa tiettyjä osa-alueita, mikä osoittaa mahdollisuuksia jatkotutkimukselle.

AVAINSANAT: business intelligence, data processing, data storage, visualisation, warranty

Contents

Figures

Tables

Abbreviations

BI	Business Intelligence
ETL	Extract Transform and Load
ELT	Extract Load and Transform
DSR	Design Science Research
DSRM	Design Science Research Methodology
IS	Information Systems
WDA	Warranty Data Analysis
OLAP	Online Analytical Processing
EDW	Enterprise Data Warehouse
AI	Artificial Intelligence

1 Introduction

Granting a warranty for a product provides assurance to customers that they will receive a product of acceptable quality (Vintr & Vintr, 2007, p. 323). A warranty defines the promised level of quality as well as the compensation available to the customer if the warranty conditions are not fulfilled. However, offering a warranty produces additional expenses for the manufacturer, commonly referred to as warranty costs. Wallace et al. (2011, p. 8) estimate that these costs typically range between 2% and 10% of the sale price. Consequently, warranty costs can have a significant influence on a manufacturer's overall profitability (Murthy, 2006, p. 134).

Accurate warranty analysis is essential for manufacturers (Wang et al., 2019, p. 1047). According to Wu (2012, p. 795) the objective of this analysis is to provide useful information and support decision-making. Such information can be utilized for multiple purposes, including identifying faulty components, supporting product development, analyzing and explaining warranty costs, predicting future claims and costs, and evaluating product reliability when determining warranty policies. However, inaccurate warranty analysis may lead to negative outcomes. For instance, overestimating or underestimating warranty costs can lead to financial losses. Overestimation may result in higher product prices and reduced sales, whereas underestimating them can cause a lack of liquid assets to cover actual expenses (Wang et al., 2019, p. 1047).

In the company involved in this research, warranty related aspects are in a digital form as data. One way of transforming this data into information is through business intelligence (BI) processes (Golfarelli et al., 2004, p. 1). These processes enable companies to collect, store, access, and examine data for achieving better business decisions related to, for instance, vendors, customers, employees, supply chain operations, and organizational infrastructure (Lawton, 2006, p. 14). Building on this view, Negash and Gray (2008, p. 176) define BI as systems that connect data extraction, storing, management, and analysis for evaluating complicated organizational and competitive information and for

presenting it to decision-makers with the objective of enhancing the quality of information utilized in the decision-making processes.

Some internal warranty data is already utilized for business intelligence purposes within the participating company in this study, but there is a need for a management-level solution that provides new insights into warranty data. In addition to internal data, external climate data is an important factor in the warranty team's decision-making, as the company operates globally and its products are significantly affected by surrounding climate. Despite its relevance, climate data has not yet been incorporated into warranty data analysis within the company. Integrating internal warranty data with external climate data in a visual format, implemented using suitable software, would improve understanding of the drivers behind warranty outcomes and enhance management's ability to make data-driven decisions across different operating regions and environmental conditions.

As the volume of available data continues to grow and its utilization increases, the company has placed greater emphasis on continuous technological development and data quality assessment. This has resulted in the adoption of modern data architectures that separate computationally intensive data processing from the visualization layer. This new approach shows potential for implementation within a modern lakehouse environment, Databricks, in combination with a business intelligence tool, Power BI. Regardless of the effectiveness of analytical tools, the value of analysis is constrained by the quality of the underlying data (King & Schwarzenbach, 2020, p. 26). If data quality is insufficient, the resulting information cannot be considered reliable.

In response to these challenges and needs, the mentioned definition of business intelligence provided by Negash and Gray (2008, p. 176) forms the foundation of the process that is addressed in this research. The objective of this research is to develop a method for building a warranty data analysis dashboard for the company involved. The method includes data gathering, processing, and visualization, with the aim of providing reliable

information to support warranty-related decision-making. To achieve these objectives, this study utilizes modern technological solutions. The workflow is intended to be implemented using Databricks for data processing and computational tasks, and Power BI for data visualization.

The prior research has examined components relevant to this study, such as warranty data analytics, extract-load-transform (ELT) and extract-transform-load (ETL) processes, lakehouses, and visualization methods, but typically as separate elements rather than as a unified system. What is lacking is an integrated approach that combines data from different sources within a lakehouse, incorporates data quality considerations, and presents the insights through a dashboard. The absence of such solution forms a clear research gap that this study seeks to address.

The methodology adopted in this study is design science research (DSR), which is applied to develop and document the proposed process. DSR focuses on solving identified organizational challenges by developing and assessing innovative IT artifacts (vom Brocke et al., 2020, p. 8; Hevner et al. 2004, p. 77). It consists of three cycles referred to as relevance, design, and rigor (Hevner, 2007, p. 88). The relevance cycle connects the research to its practical context, the design cycle concentrates on developing and evaluating the artifact, and the rigor cycle grounds the work in existing scientific knowledge. The research is implemented using the steps of design science research methodology (DSRM) provided by Peffers et al. (2007). The artifact chosen for this study is a method, which, according to Winter (2008, p. 471), is a structured process that provides instructions for solving a certain problem.

The contribution of the method lies in its ability to guide organizations in developing visual analytics solutions that integrate different data sources while ensuring data quality and leveraging new technologies. Unlike a single software implementation, the method is transferable and can be applied in similar organizational contexts. Although this study focuses on warranty and climate data, the data itself serves primarily as an example. The

fundamental mechanisms of the method remain applicable regardless of the type of data. Another contribution of the method, specifically within the context of warranty management, is its ability to reveal high-quality, trustworthy insights that would otherwise remain hidden.

The research questions are developed based on the guidelines proposed by Thuan et al. (2019, p. 349). According to these guidelines, research questions in DSR can be aligned with the relevance, design, and rigor cycles, and concrete examples are provided to support their formulation. However, Thuan et al. also emphasize how DSR may implement part of the instructions, combine the templates, or add other questions. Accordingly, the research questions are derived from the templates to suit the goals in this research while ensuring that each cycle is addressed.

The exact research questions are:

1) What prior knowledge is available to inform the design of the method artifact?
2) What steps should the method include based on prior knowledge and input from company domain experts?
3) To what extent is the developed method applicable for addressing the identified problem in practice?

As reflected in the research questions, this study first examines prior knowledge related to the components of the method. This knowledge, together with insights from the company domain experts, forms the foundation for designing the method. The method then specifies the steps required to develop the solution. Finally, the demonstration of the method and the resulting dashboard addresses whether the method is applicable for solving the identified problem in practice.

The research is structured as follows. Chapter 2 presents the knowledge base through a literature review covering warranty, data integration, data storage, and business

intelligence, and identifies the research gap. Chapter 3 introduces the research methodology, explaining the principles of DSR and how it is applied in this study, as well as the research environment and how data is gathered. Chapter 4 contains the resulting artifact in the form of a method, documenting the steps of the design and development process in a way that can be generalized for future use. This chapter also presents the demonstration and the results obtained from it. Chapter 5 concludes the study with final remarks.

2 Literature review

This chapter presents the essential theoretical background required to understand the key concepts of the study and to identify components that can be used in the proposed method. The concepts of warranty, data integration, business intelligence, and data storage are reviewed. The chapter concludes by identifying the research gap that this study seeks to address.

2.1 Warranty

A warranty is a commitment by the seller that the product will perform as promised (Blischke & Murthy, 1995, p. 6). Jack and Van Der Duyn Schouten (2000, p. 95) expand this definition by describing a warranty as a contract between the seller and the buyer that obligates the seller to repair any failure or replace the malfunctioning product within the warranty period. The decision to repair or replace the product is determined by the manufacturer, and depends on the associated costs, the product's lifespan, and how close to the end of the warranty period the failure occurs.

From the buyer's perspective, the primary function of a warranty is to protect the right to seek compensation when a properly used product fails to function as promised or as specified (Blischke & Murthy, 1995, p. 8). Additionally, a buyer's perception of a product can vary depending on the length of its warranty. Products with longer warranty periods are often viewed as more reliable and durable, whereas shorter warranties may lead consumers to perceive the product as lower in quality. From the seller's perspective, a warranty outlines the intended use of the product, specifies the required maintenance, and clarifies the circumstances under which the seller is not obligated to provide compensation. In addition, warranties have an important marketing function for the seller, as a longer warranty can be used to attract customers and create a competitive advantage over other competitors (Blischke & Murthy, 1995, p. 9).

2.1.1 Warranty costs

Warranty obligations create additional costs for manufacturers referred to as warranty costs (Wallace et al., 2011, pp. 8–9). Typically, warranty costs account for approximately 2-10 % of a product's sales price, making them a relatively significant expense. Several factors influence warranty costs, including the terms of warranty, customer usage patterns, and product reliability. Warranty terms are defined by warranty legislation and are further shaped by how competitors structure their own warranties. Customer usage, however, varies widely and remains outside the manufacturer's control. Consequently, product reliability becomes the only factor that the manufacturer can actively influence, but it has a significant impact on warranty costs. Therefore, one approach to reducing warranty costs is to enhance product reliability. However, such improvements often require substantial investment in research and development. These investments are justified if the benefits of the development expenses exceed the corresponding warranty costs.

2.1.2 Warranty claims

Warranty claims occur when a product or its components fail under the warranty period (Wallace et al., 2011, p. 61). Wallace et al. (2011, p. 64) define failure as "inability of an item to function as required when operated properly". However, not all failures lead to a submitted warranty claim. Customers may be dissatisfied with the claim process, may switch to another brand, may perceive the effort required to submit a claim as excessive, or may simply forget that the product is still under warranty. Once a claim is initiated, the warranty servicing process begins.

The first step of the warranty servicing process typically involves collecting data and information about the failed product or component (Wallace et al., 2011, p. 64). After this, warranty personnel evaluate whether the claim is valid. Several factors may lead to a rejection of claim such as expired warranty period, lack of evidence that the item has failed, the product being resold, or if the item was not used in accordance with the

warranty terms. If the claim is accepted, the next stage is the technical resolution of the issue. In cases where a replacement is provided, the failed component is usually sent back to the manufacturer for further analysis to determine the root cause of the failure. A company that collects warranty data as part of the aforementioned process can leverage it to gain valuable insights into the failures occurring during the warranty period (Araujo, 2023, p. 1).

2.1.3 Warranty data analysis

The literature identifies several categories of warranty data that are relevant for analytical purposes. According to Wallace et al. (2011, p. 10), warranty data contains the information required for the effective management of warranties for both existing and newly introduced products. This data can be divided into groups of warranty claim data and supplementary data. Warranty claim data is gathered through support systems during the processing of warranty claims and repairs under warranty, typically after the sale. Supplementary data, in contrast, originates from additional internal or external sources within or outside the manufacturing organization.

Supplementary data can be further categorized into several groups. Product-related data include technical details such as failure of a component, fault types, product age, and usage at the time of failure (Wallace et al., 2011, p. 10). Customer-related data describe aspects such as state of usage, intensity of use, operating environment, and maintenance behavior. Maintenance service data contains maintenance actions performed by the manufacturer, the quality of service and repair expenses. Market-related data, in turn, includes the performance of competing products, pricing, and warranty conditions.

Furthermore, Wu (2013) examined various types of coarse warranty data and analytical approaches, identifying several types of data relevant for warranty data analysis (WDA). The claim-related data identified in the study included, for instance, the number of products with one or more warranty claims, the number of products with claims occurring within specific sales or manufacturing time frames, and the number of products with no

claims (Wu, 2013, p. 2). In addition, warranty datasets may also contain date-related data, such as the day on which a warranty claim was conducted, the day a failure occurred, the day the item was sold, and the day the item was produced.

Marshall et al. (2018, p. 544) note that warranty cost related data can be dependent on frequency of claims, repair time and the claim's cost, while Blischke (1993, p. 84) highlights the importance of analyzing warranty costs per unit sale and lifetime operating costs of a unit. Annadurai (2023, p. 1) further emphasizes how senior management often focuses on information about products with high and low warranty cost as well as total warranty costs broken down by sales year, region, supplier, and customer.

Beyond understanding the types of warranty data, the quality and management of data are critical for producing accurate analysis. Wu (2013, p. 2) notes that warranty data may be incomplete, summarized, delayed, censored, or imprecise. Although the variables presented in their study can be used to conduct warranty data analysis, Wu emphasizes that achieving reliable results requires careful attention to data quality. Wallace et al. (2011, p. 10) similarly stress the importance of proper data collection, noting that inaccurate gathering of warranty data can cause significant loss of information and may have a negative effect on data-driven decision-making. A practical example of the impact of data limitations is presented by Mahlamäki et al. (2016), who developed a model for predicting warranty costs. Although data were collected from multiple installations, a substantial portion had to be discarded due to inconsistencies and missing values (Mahlamäki et al., 2016, p. 5). The data cleaning process proved to be burdensome, which caused the objectives of the study to be compromised. Their study clearly illustrates how the reliability of analytical outcomes depends on the availability of high-quality and consistent data.

An example of the process of analyzing warranty data is presented by Annadurai (2023) who discusses how WDA can be automated using data science techniques. According to the paper, automating WDA provides several organizational benefits, including the

ability to conduct analyses with limited prior experience, access to valuable reliability metrics, enhanced data visualization, the possibility to examine large product populations simultaneously, and improved filtering options for analysis (Annadurai, 2023, p. 1). The author explains that WDA typically involves extracting the required data from databases, transforming raw data to meet the intended analytical purpose, and performing analysis to obtain insights. Their method for analyzing product life data included grouping products according to their risk levels, preparing the data by estimating usage from failure information, assigning products to the appropriate groups, and conducting data analysis supported by visualizations (Annadurai, 2023, pp. 2–4).

2.2 Data integration

Data integration refers to the process of combining data from different sources into a unified view (Sherman, 2014, p. 272). Sufficient data integration should support the entire organization rather than a single project (Sherman, 2014, p. 274). It should proceed gradually in manageable steps, allowing learning from each integration phase. According to Walha et al. (2024, p. 26688), it typically includes the use of extract, transform, and load (ETL) workflows, which gather data and convert it into a consistent format. To accelerate data preparation process, extract, load, and transform (ELT) has also emerged as an alternative (Jo & Lee, 2019, p. 6). By providing relevant information, the data integration process enhances the organization's ability to make informed decisions (Walha et al., 2024, p. 26688). In addition, according to Souibgui et al. (2019, p. 677), data quality is a crucial component of the ETL process, and therefore also needs to be addressed.

2.2.1 Data quality

Data is an organization's most valuable asset, carrying real and measurable worth (Mahanti, 2019, pp. 24–26). However, the value does not only come from data itself but also depends on the actions of a company when utilizing data. Storing large amounts of data alone is not enough as the real benefits arise only when data is in appropriate quality (King & Schwarzenbach, 2020, p. 27). If existing data quality is perceived as poor, no

matter how effective the software tools are, the outcomes will be constrained by the low-quality data (King & Schwarzenbach, 2020, p. 26).

From a business intelligence perspective, the objective is to generate high-quality information that supports managerial decision-making (Wieder & Ossimitz, 2015, p. 1165). Achieving this typically involves two key stages. The first concerns identifying, gathering, storing, and managing data in appropriate repositories. The second involves retrieving, analyzing, and presenting data in a form that is useful for decision makers. Central to this process is the quality of the underlying data. According to Wieder and Ossimitz, data quality is a prerequisite, but not a guarantee, for high information quality. Nevertheless, when data is both high quality and well maintained, the resulting information is likely to be reliable.

Wang and Strong (1996) developed a framework for conceptualizing data quality, addressing the need to broaden the previously narrow focus on accuracy. In their study, the definition of data quality was expanded to include the dimensions of intrinsic, contextual, representational, and accessibility. Intrinsic data quality reflects characteristics inherent to the data, regardless of the context or task, including dimensions of correctness, objectivity, believability and credibility (Wang & Strong, 1996, pp. 19–20). Contextual data quality, in turn, focuses on the requirements that data must fulfil to be suitable for the particular task and context in which it is used. To provide contextual value, data should be relevant, current, complete, and appropriate in quantity. Representational data quality concerns how data is presented within a system. Data must be represented consistently, concisely, and in a form that is easy to interpret to ensure meaningful results. Finally, accessibility is presented as a distinct dimension rather than as a previously assumed characteristic. It requires that data is available when needed while also being adequately protected through proper security measures.

In addition to Wang and Strong's (1996) framework on data quality, more recent research by Miller et al. (2024) presents four more attributes. Using the terminology of

Miller et al. (2024, p. 6), the attributes are referred to as governance, usefulness, quantity, and semantics. Governance is related to how data follows formalized frameworks of authority and responsibility. It is driven by the organization's internal structures, guidelines, and workflows of data management, which ensure that data supports organizational functions and aligns with its values (Miller et al., 2024, p. 13). Usefulness describes the ability of data to fulfil the specific needs of its users and applications. This includes its adaptability across different contexts and its potential for reusability (Miller et al., 2024, p. 6). Quantity concerns whether the amount of available data is sufficient and provides adequate coverage for a detailed description of the information (Miller et al., 2024, p. 17). It considers the level of detail, granularity and scope required for reliable decision-making. The last attribute mentioned by Miller et al. (2024, p. 18) is semantics which refers to how well data provides the detail needed to communicate its meaning to users and applications. Semantic information includes background, descriptions, and related characteristics that assist in understanding the meaning of data across different use cases.

Based on the foundations of data quality presented by Wang and Strong (1996) and Miller et al. (2024), it is evident that data quality can be assessed through multiple perspectives. Wang and Strong (1996, p. 6) provide a foundation for what data quality means to data consumers, while Miller et al. (2024, p. 3) extend this to a framework that enables different domains to discuss data quality issues. Together, these frameworks illustrate that data quality is not a single attribute, but a scope of different properties that ensure reliable organizational analysis. However, ensuring data quality in practice relies largely on the ETL processes, making them a critical foundation for data-driven analytics (Souibgui et al., 2019, p. 677).

2.2.2 ETL process

ETL is a pipeline of extracting, transforming, and loading data (Doan et al., 2012, pp. 251–252). In this process, data is collected from the source systems, processed, and then stored in a data warehouse (Sherman, 2014, pp. 37–38). The first phase involves

extracting data from original sources, such as old systems or databases (Doan et al., 2012, pp. 251–252). In the second phase, the extracted data is transformed into a suitable format for the data warehouse. The third phase consists of loading the data into the target warehouse, either by overwriting existing data or by adding new records. These phases are discussed in more detail in the following subsections.

2.2.2.1 Data extraction

As mentioned, the first phase of ETL is extraction, during which data is collected from sources, such as operational or external systems, to a staging area of the data warehouse (El-Sappagh et al., 2011, pp. 1–2; Vassiliadis, 2009, p. 2; Simitsis et al., 2023, p. 1). Each source has its own characteristics that must be considered to ensure data is extracted adequately for the ETL process (El-Sappagh et al., 2011, p. 2). The extraction process must also integrate systems operating on diverse platforms, for instance operating systems, database management systems, and multiple communication protocols. Khan et al. (2024, p. 3) mention three extraction methods that are extracting the entire dataset from the source system each time, extracting only a part of the dataset without notifying about the update, and extracting only a part of the dataset with notifications that indicate updates. Regardless of the method, extraction should not have an impact on the performance of the source system.

2.2.2.2 Data transformation

During the transformation stage, the extracted data is modified and processed through a series of functions to ensure it becomes, for instance, accurate, consistent, and precise. (El-Sappagh et al., 2011, p. 3; Khan et al., 2024, p. 4). This stage typically involves cleaning and conforming data to make it accurate for future use (El-Sappagh et al., 2011, p. 3). According to Borrohou et al. (2025, p. 3) there are several different data transformation techniques that exist, each designed to convert data into different formats and to support a variety of use cases. The choice of transformation technique depends on the intended purpose, as well as on the file formats and data structures involved.

Ponniah (2010) provides the key data transformation functions that form the core tasks of data transformation, regardless of the complexity of source systems. The first function, selection, involves choosing the relevant parts of records from the source systems (Ponniah, 2010, p. 296). Although selection is typically associated with the extraction phase, certain source system structures do not allow partial selection. In such cases, the complete record is extracted, and selection is performed during the transformation stage. Another form of data transformation includes splitting and joining which allows the selected data to be divided into separate components or combined by joining tables. Conversion refers to standardizing data extracted from different sources, as well as improving the clarity and usability of the fields for users. Summarization aggregates detailed data into a higher level of granularity, reducing unnecessary detail when it is not needed (Ponniah, 2010, p. 297). The final transformation identified is enrichment, which enhances data usefulness by organizing and simplifying individual fields. Enrichment is applied when one or more fields are derived from multiple records, thereby creating a single field for that data.

2.2.2.3 Data loading

Loading data into the target system represents the final phase of the ETL process. Loading is typically performed in the background at predetermined intervals (Biswas et al., 2019, p. 57). According to Khan et al. (2024, p. 5), there are three main approaches to data loading. The first is the initial load, in which all tables in the data storage are populated for the first time. The second is the incremental load, where data is imported at scheduled intervals to implement continuous updates. The third is the full refresh, in which the contents of one or more tables are completely deleted and then reloaded with new data. However, Biswas et al. (2019, p. 57) explain that incremental loading is more efficient than full refresh, as it identifies the records that have changed in the source data and sends only those updates to the database.

2.2.3 ETL and ELT

ETL is typically a complicated combination of processes and technologies, and it accounts for a considerable portion of data warehouse development while being costly, complex and time consuming (El-Sappagh et al., 2011, p. 2; Walha et al., 2024, p. 26688). According to El-Sappagh et al. (2011, p. 2), ETL development also demands expertise from a range of specialists, such as business analysts, database designers, and software developers. It is a continuous process, since data sources change over time, and the data warehouse requires regular updates to remain accurate. In addition, as business requirements evolve, the data warehouse must be adapted to maintain its value as a decision-making tool. Therefore, ETL processes should be designed to be easily modifiable and well documented.

ELT has emerged as a preferred approach when data volumes continue to grow (Storey & Song, 2017, p. 56). ELT stands for extract, load, and transform, indicating that data is first extracted and loaded into the target destination, after which the transformation operations are executed within that environment (Jo & Lee, 2019, p. 6). The comparison between ETL and ELT is depicted in Figure 1. According to Simitsis et al. (2023, p. 4), in many cases the extract and load phases function as data replication, and the challenge is to implement it as efficiently, securely, and precisely as possible. Furthermore, Dmitriyev et al. (2015, p. 53) argue that ELT is generally the preferable approach for business intelligence (BI) due to its flexibility and adaptability.

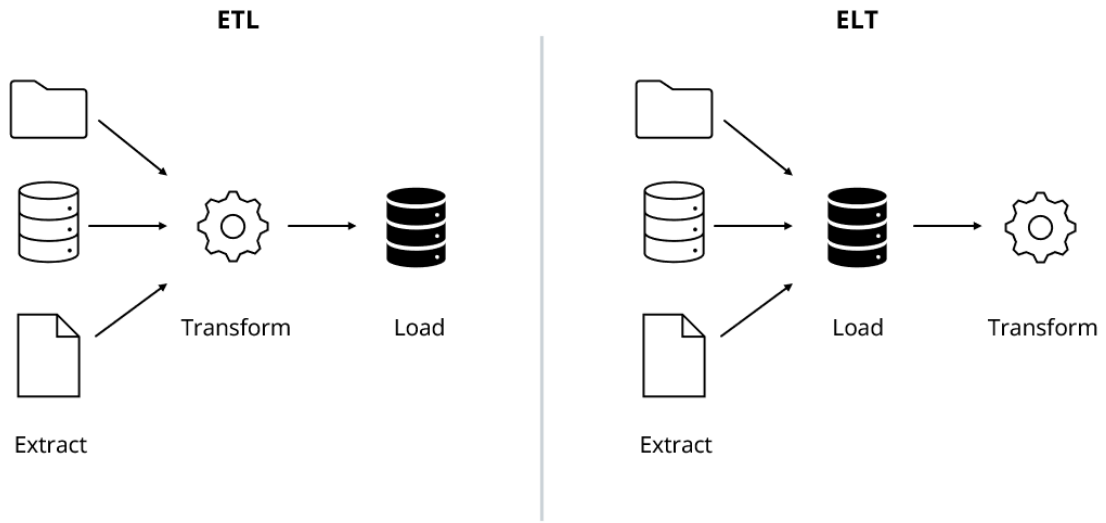

Figure 1. Comparison of ETL and ELT.

The growing popularity of ELT can be explained by several reasons. First, ELT is suitable for environments where data is generated in continuously larger volumes, often without human intervention (Simitsis et al., 2023, p. 4). Contemporary cloud data platforms further support this shift by enabling cost-effective analysis of remote and distributed data sources in a centralized environment. At the same time, computing costs have decreased due to the rise of open-source technologies such as Apache Spark and the availability of cloud services including Microsoft Azure. According to Dmitriyev et al. (2015, p. 53), ELT enables a more flexible addition of new data sources compared to ETL, as data is transformed only within the target system. This approach also allows raw data to be reprocessed multiple times as needed, whereas ETL requires data to be transformed before loading it into the target database. Furthermore, in ELT, transformations are executed only when necessary, which is crucial for rapidly evolving business models. Dmitriyev et al. conclude that these characteristics make ELT preferable for BI solutions.

2.3 Business intelligence

According to Sherman (2014, pp. 21–22) knowledge is considered a vital component of a successful enterprise. Knowledge is derived from information, which in turn originates from data. Business intelligence (BI) enables this transformation by converting data into

meaningful knowledge that supports organizations in making informed decisions across all operational stages. Khan et al. (2020, p. 116013) define BI as a set of methods, techniques, and concepts that positively influence business decisions through data-driven systems. BI not only enhances performance and effectiveness but also contributes to cost reduction and loss minimization. Consequently, BI provides strategic guidance by enabling executives to identify patterns that might otherwise remain unnoticed (Sherman, 2014, p. 25). This is crucial, as data holds limited value unless it is properly understood, analyzed, and utilized (Sherman, 2014, p. 23).

BI continues to experience increasing demand among organizations (Sherman, 2014, p. 27). Jiménez-Partearroyo and Medina-López (2024, p. 19) argue that BI functions not only as a tool for understanding information but also as a strategic asset that shapes and adapts organizational decision-making. Moreover, BI has evolved from a specialized tool into an integrated part of business strategy and is now increasingly applied for new purposes. Additionally, BI enhances an organization's competitive differentiation by enabling rapid adaptation and trend prediction (Jiménez-Partearroyo & Medina-López, 2024, p. 19; Sherman, 2014, p. 27). The following subsections examine the central business intelligence concepts relevant to this research.

2.3.1 Data modeling

Data modeling is one of the most important features in BI implementation (Sinha, 2021, p. 82). A data model defines and maintains how data is structured, making it essential for the effective use of data (Ballard et al., 2006, p. 25). The most widely used data model is the dimensional model (Sinha, 2021, p. 5). It is composed of two types of tables referred to as dimension and fact tables. Dimension tables store descriptive or qualitative attributes that provide context for the fact tables. Each dimension table contains a primary key, which corresponds to a related foreign key in the fact table. The fact table stores measurable or quantitative attributes of business data (Sinha, 2021, p. 7). As illustrated in Figure 2, the combination of a single fact table with multiple surrounding dimension tables forms a star schema (Sinha, 2021, p. 6). When dimension tables are

further normalized into separate tables that link back to the main dimension table, the resulting structure is referred to as a snowflake schema, as depicted in Figure 3 (Ballard et al., 2006, p. 74). However, Ballard et al. (2006, p. 165) do not recommend the snowflake schema, as it can cause slower performance due to increased joins between tables and may also reduce the understandability of the data model.

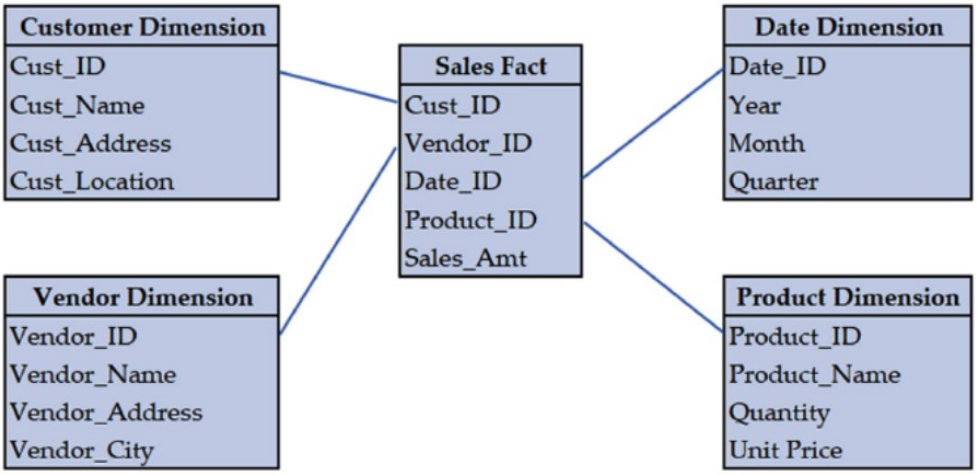

Figure 2. Star schema (Sinha, 2021, p. 8).

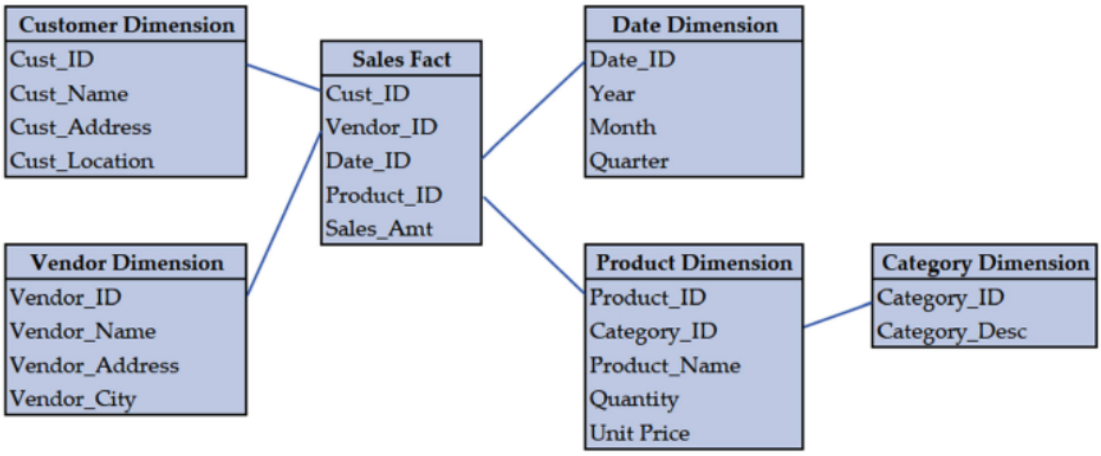

Figure 3. Snowflake schema (Sinha, 2021, p. 9).

2.3.2 Dashboard implementation

BI is a form of descriptive analytics within the broader field of business analytics, focusing on building reports which answer questions “What happened?” and “What is happening?” (Sharda et al., 2018, p. 130). These reports provide decision makers with records of business transactions at regular intervals, the ability to generate customized views, and access to key performance indicators through graphical interfaces. To produce these graphical interfaces referred to as dashboards of information, appropriate selection of charts is required (Sherman, 2014, p. 343). For instance, in a study by Muntean et al. (2021, p. 13), visualization types were chosen according to the analytical objectives and required insights. Additionally, in a study by Silva et al. (2020, p. 12) the charts were selected during the prototyping of a dashboard. Sherman (2014) further discusses how the choice of visualizations should be guided by the type of analysis, which is examined in the following paragraph.

For comparative analysis, where the objective is to compare or rank data, bar charts are efficient as the heights of bars depict measurable values, enabling direct comparison (Sherman, 2014, p. 343; Saket et al., 2019, p. 7). Time-series analysis focuses on changes over time and typically utilizes bar or line charts (Sherman, 2014, p. 344). Bar charts suit situations with few segments or when continuity is unimportant, whereas line charts emphasize also the trend. Contribution analysis, which illustrates how segments contribute in percentages, often uses pie charts (Sherman, 2014, pp. 344–345). However, their effectiveness decreases with more than four segments, in which case bar charts are preferable. With larger number of segments, heat maps with informative color gradients are more effective. Correlation analysis, which seeks to identify relationships between variables, is most effectively visualized using scatter plots. Distribution analysis, which illustrates how values spread across a range, is effectively visualized with box plots or histograms. Finally, location-based data is typically visualized using maps.

Saket et al. (2019), Sedrakyan et al. (2019), and Srinivasan et al. (2019) provide empirical evidence on chart usage that align with Sherman’s recommendations. Saket et al. (2019,

p. 2511) who evaluated the effectiveness of different visualization types, found that scatter plots and line charts provide higher accuracy and faster performance in correlation analysis. They further reported that bar charts enable accurate and efficient comparative analysis. Sedrakyan et al. (2019, p. 28) demonstrated that area and line charts can highlight changes over time and enhance the visibility of trends. Their study also showed that pie charts are suitable for contribution analysis, referred to in their study as part-to-whole analysis. They also noted that map visualizations are the most straightforward option for representing location-based data (Sedrakyan et al., 2019, pp. 32–33). Srinivasan et al. (2019, p. 673) recommended the use of box plots and histograms for distribution analysis, consistent with Sherman's theoretical suggestions.

Studies have also examined the layout of dashboard charts. According to the research by Zhang et al. (2024, p. 19), who investigated the visual order and the placement of core charts using eye tracking experiments, the central challenge in dashboard design is to reduce the negative effects of complexity while preserving the quantity of information presented. It was found that the optimal layout of dashboard is to locate the core chart to the left center position, while the remaining non-core charts are arranged in a partially symmetrical structure (Zhang et al., 2024, p. 19). This arrangement mitigated the negative effects of dashboard complexity, and lead to faster visual search performance (Zhang et al., 2024, pp. 17–18).

Different design principles are also valuable in the development of dashboards. Research by Ivanković et al. (2021) examined dashboards presenting Covid-19 data to derive design principles for dashboard design that support decision-making. While their study focused on dashboards presenting Covid-19 data, several of the identified principles are applicable to dashboard design more broadly. The first principle emphasizes the importance of clearly defining the target audience (Ivanković et al., 2021, p. 12). Dashboards that are designed with a specific user group in mind tend to have more coherent content, analysis, and visual displays. The second principle highlights the value of using a moderate number of indicators. Limiting the use of indicators emphasizes the

importance of information. Additionally, structuring the information by moving from general to detailed, or grouping it by theme, creates an intuitive flow of information. Finally, the third principle relevant to overall dashboard design stresses the importance of making data sources and methods transparent. Providing clear explanations of how indicators have been constructed, as well as mentioning their limitations, was found to enhance trust in the dashboard.

Furthermore, Bach et al. (2023) analyzed different dashboard types and proposed design guidance to support dashboard development. According to their model, presented in Figure 4, the goal of the design process is to balance four key parameters of abstraction (amount of information), screen space, number of pages, and interactivity (Bach et al., 2023, p. 347). These parameters are in tension with each other, and achieving an ideal balance requires design trade-offs. This means that an increase of one parameter necessitates a decrease in one or more of the others, and vice versa. The model illustrates this interdependence by showing how adjustments to a single parameter influence the remaining ones.

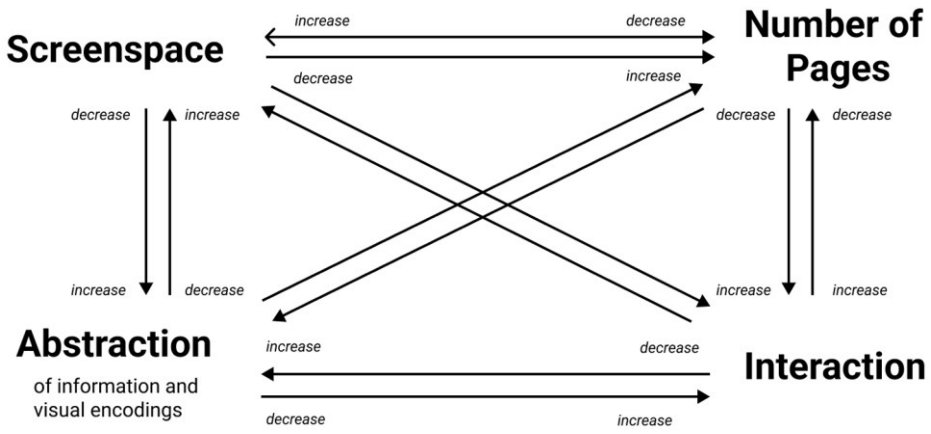

Figure 4. A simplified model for design trade-offs in dashboard design (Bach et al., 2023, p. 347).

For instance, according to the model, increasing screen space can reduce information loss, but large displays cannot always be assumed, as users may access dashboards on mobile devices (Bach et al., 2023, p. 347). When a high amount of information must fit

on a regular size screen, careful attention to page layout is required, such as deciding how many pages to use, what content appears on each page, and how components are arranged. Dashboards may also be designed as simple and static one-page displays that present information concisely without interactivity, but such approach requires careful component placement to ensure efficient use of the available screen space. Interactive dashboards, in contrast, provide middle ground between static and paginated dashboard approaches (Bach et al., 2023, p. 348). Visible information may be minimized to fit the screen while allowing users to choose what additional information to access. Interactivity can therefore control both the amount of information and the pace of the presentation, for instance by revealing information gradually based on the user's actions. This can be enabled through navigation or scroll buttons, tabs, links, or detail-on-demand features, where data remains hidden until triggered by user interaction.

2.3.3 Power BI

Power BI is a business intelligence application that provides tools for analyzing and visualizing data (Hashemi-Pour et al., 2024). It supports users with varying levels of technical expertise, including business users, report creators, and developers (Microsoft, 2026). Power BI enables data collection from multiple sources, transformation of data to meet business requirements, and presentation of information through graphs, maps, and dashboards (Sinha, 2021, p. 11). Sinha (2021, p. 12) identifies the core components of Power BI, which are discussed in the following section.

Power BI Desktop is the primary interface within the Power BI environment, enabling users to connect and transform data, create calculations, and design reports (Sinha, 2021, p. 12). Data preparation is performed using Power Query, which allows various transformations to be applied to tables (Sinha, 2021, p. 13). These transformations are executed using the M programming language in the background. Power Pivot functions as the calculation engine in Power BI and is used for defining relationships between tables and implementing calculations using the data analysis expressions language (Sinha, 2021, p. 14). Visualizations are created with Power View, which enables the development of

tables, charts, and maps. Its drag-and-drop interface facilitates rapid construction of visual elements. Lastly, Power BI Service provides a cloud-based environment where reports created in Power BI Desktop can be published. It supports sharing visualizations across the organization and allows data to be refreshed automatically.

2.4 Data Storage

Storing data originally evolved to support business intelligence by moving operational data into data warehouses, which are optimized for analytical workloads (Armbrust et al., 2021, p. 1). As data volumes increased and more information became unstructured, data warehouses proved too costly and incapable of storing such data. This led to the emergence of new platforms called data lakes which store raw data in inexpensive and flexible storage systems using open file formats and a schema-on-read approach. In many current architectures, data warehouses are utilized for storing data while data lakes handle computing. However, maintaining two separate systems and ETL pipelines increases complexity. The data lakehouse has emerged as a solution that integrates the strengths of both data warehouses and data lakes (Armbrust et al., 2021, p. 2). The following subsections discuss these concepts in more detail to provide the necessary background.

2.4.1 Data warehouse

A data warehouse is an integrated and historical data repository (Gaol et al., 2023, p. 2). It generally supports strategic decision-making through techniques based on online analytical processing (OLAP), in which data is organized either in relational databases or in multidimensional structures (Gaol et al., 2023, p. 2; Chaudhuri & Dayal, 1997, p. 2). Data is typically gathered from independent and diverse sources using processes such as ETL and combined into a unified, well-structured view (Chandra & Gupta, 2018, p. 1; Gaol et al., 2023, p. 2). Gaol et al. (2023, p. 2) further describe a data warehouse as a centralized database that acts as a data center, created by connecting and managing data from different sources. Organized to support efficient queries, reporting, and analysis, a data

warehouse often functions as a foundation for decision-making (Al Olimat et al., 2025, p. 65).

According to Sherman (2014, p. 30), data warehouses provide greater value for an organization's business intelligence strategy than connecting data visualization tools directly to operational data sources that capture transactional data. While operational data structures are designed to support business transactions and communications, data warehouse structures are optimized specifically for analytical purposes. Organizations require the ability to perform period-over-period comparisons and inspect trends using historical data that operational systems do not support. Furthermore, operational data is often distributed across multiple source systems, which complicates integration and analysis. Lastly, organizations rely on both company-level and business-unit-specific performance indicators (KPIs), which typically must be derived and calculated outside of operational systems.

2.4.2 Data lake

Fang (2015, p. 820) defines a data lake as a methodology that relies on a large-scale, low-cost data repository to improve how enterprises collect, refine, store, and analyze raw data. A data lake can store unstructured or multi-structured raw data that mostly holds unknown value for the organization. It has gained popularity because it provides a cost-efficient and technologically functional approach to address the challenges of big data.

A data lake enables several functionalities. Large volumes of raw data can be collected and stored in a data lake at a low cost, which is an important factor as data volumes continue to grow (Fang, 2015, p. 820). Data lakes also support processing of various data types, including structured data from traditional databases, multi-structured data with undefined attributes, and multimedia data such as text, graphs, or videos. Additionally, data in a data lake can be preprocessed and transformed, enabling other systems to explore the prepared data. Furthermore, data lakes utilize a schema-on-read approach, where the structure of the data is determined at the moment of use, eliminating the

need for complex data modeling and integration efforts. Finally, data lakes enable analytics for specific use cases, which is beneficial when the value of the data is still uncertain. This allows organizations to explore the data and determine how it can be utilized.

Generally, data lakes lack standard data warehouse functionalities, such as data quality controls, transactional support, or mechanisms of maintaining consistency (Simitsis et al., 2023, p. 5). Therefore, enterprises often incorporate a combination of a data lake and a data warehouse into their architecture. It involves moving data to data lakes using ETL processes, followed by transferring it into data warehouses using ELT (Armbrust et al., 2021, p. 1). However, this integration introduces several challenges. Storage requirements increase due to storing multiple copies of the data, infrastructure and operational costs rise, previous analytical results can be difficult to trace, and maintaining security and access control can raise concerns (Simitsis et al., 2023, p. 5).

2.4.3 Data lakehouse

Armbrust et al. (2021, p. 1) argue that data warehouse architecture is likely to undergo significant changes in the near future and may be replaced by a new approach known as the lakehouse. Figure 5 illustrates the lakehouse model, showing how structured, semi-structured, and unstructured data can be processed within the lakehouse. It also shows how this architecture extends the traditional data lake by incorporating performance features of data warehouses, thereby combining the strengths of both systems and enabling a wide range of use cases, including business intelligence. According to Armbrust et al. (2021, p. 2) this combination is enabled by Delta Lake, which significantly improves how data can be managed and how it performs in cloud object storage. The authors also highlight that industry implementations show promising outcomes, suggesting that the

lakehouse architecture can resolve the main challenges with data warehouses and data lakes.

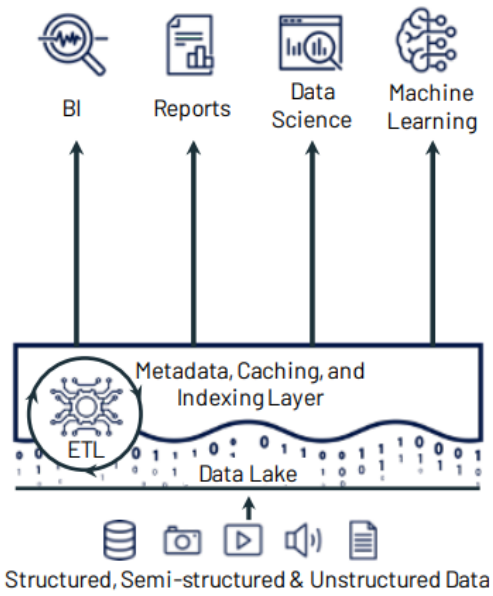

Figure 5. The lakehouse model (Armbrust et al., 2021, p. 2).

Armbrust et al. (2021, p. 2) describe various solutions and functionalities provided by a lakehouse. It supports flexible storing of raw data in a manner similar to a traditional data lake, allowing data to be accessed at any time, while also incorporating the advanced management and performance-optimization capabilities of data warehouses (Armbrust et al., 2021, pp. 2–3). This architecture enables ETL and ELT processes to improve data quality and reduces the number of steps organizations must perform. Moreover, the lakehouse provides capabilities for machine learning and data analytics while benefiting from its integrated query-optimization features. Schneider et al. (2024, p. 16) explain that the lakehouse facilitates OLAP workloads by enabling interactive querying and analytical operations on datasets stored within the lakehouse, employing query languages similar to the use of relational database tables. This means that analysts should be able to query any stored datasets for analyzing data (Schneider et al., 2024, p. 22).

The computing performance of a lakehouse must be excellent to process massive datasets stored in open formats such as Parquet or ORC (Armbrust et al., 2021, p. 2). Since these formats are openly accessible, various applications can read the data directly from the lakehouse. Consequently, the system can expose its data through public APIs, enabling faster data access compared to traditional data warehouses (Armbrust et al. 2021, p. 5). Schneider et al. (2024, p. 16) further note that, in order for the lakehouse to provide these capabilities, structured, semi-structured, and unstructured data must be stored in a unified storage system and follow a consistent data format. In their comparative analysis of data warehouses, data lakes, and the lakehouse model, they found that relying on a single storage system and a unified format simplifies the overall architecture by reducing the number of data movement and transformation steps. This simplification also eliminates the need for duplicating data across multiple systems, thereby enabling a single point of data access.

The data lakehouse has strong layers of data quality checks and validation, which are first performed within the ingestion and staging areas (Sundar et al., 2022, p. 8). At the ingestion layer, the base level structural and technical validations are conducted to ensure that the files are intact, follow the expected schema, and contain all required fields. Rows that fail these validations are redirected to designated queues for further inspection. More complex validation rules such as checks for range, anomaly detection, and removing duplicates are executed in the transformation layers. The lakehouse further enhances trust in curated tables by implementing automatic data validation and following data quality objectives such as freshness and completeness.

Sundar et al. (2022, p. 8) note that while the traditional ETL processes remain applicable within the lakehouse, ELT approaches complement them by enabling transformation activities to be implemented within the lakehouse environment. Their research explains that contemporary lakehouse teams adopt a hybrid model, in which only minimal ETL processing is performed at the “edge”, such as critical validations, format changes, and the handling of personally identifiable information. The remaining business logic,

aggregations, and data modelling are executed using the ELT model directly within the lakehouse. This shift reduces the complexity of data pipelines, accelerates the onboarding of new data sources, and enables teams to move transformation tasks closer to the environment where analytics are performed. Furthermore, it supports the delivery of raw, standardized, and data-product-ready views without the need for repeated data movement.

2.4.4 Databricks

Databricks provides the capabilities required to support the previously described lakehouse functions. It is built on a lakehouse architecture that integrates data lake and data warehouse principles (Databricks, 2026a; Rayarao & Donikena, 2025, p. 2). The platform supports data pipeline processes, data warehousing workloads, and generative AI-driven analytics (Databricks, 2026a). Data quality is managed through a medallion architecture, which organizes data into layers (Databricks, 2025). The bronze layer stores raw data, the silver layer contains cleaned and validated data, and the gold layer provides enriched data with aggregations and dimensional models. Databricks also provides natural language assistance for attempts to write code, fix errors, and find answers (Databricks, 2026a). Business intelligence tools such as Power BI can establish a direct connection to Databricks compute clusters, allowing users to perform analytics on large-scale data independently (Rayarao & Donikena, 2025, p. 7).

Databricks supports collaboration in multiple ways across an organization. It allows different teams to access shared data, AI models, and analytics while avoiding unnecessary data replication (Gaur et al., 2025). For instance, metric views, used to calculate key performance indicators, can be defined once and shared with authorized users (Databricks, 2026b). In addition, Databricks notebooks provide a shared workspace where users can collaborate in real time using multiple programming languages, including Python, R, SQL, and Scala (Rayarao & Donikena, 2025, p. 3). By supporting multiple programming languages, the platform enables individuals with different technical backgrounds to collaborate seamlessly within shared projects.

2.5 Summary of components

The literature review identified several components that can guide the development of the method and provide a theoretical basis for the design decisions. To identify how these components can be utilized within the method, Table 1 summarizes those applicable without modification, whereas Table 2 presents those requiring adaptation. The tables also include references and explanations for why each component can be applied. This approach clarifies which components can be directly utilized and which require further refinement to suit the method.

Table 1. Components applicable to the method without modification.

Component	**Explanation**	**References**
The emphasis on warranty data quality	The emphasis provides justification on why warranty data quality must be taken care of.	Wu, 2013; Wallace et al, 2011 and Mahlamäki et al., 2016
The emphasis on overall data quality	The emphasis provides justification on why overall data quality must be taken care of.	Mahanti, 2019; King & Schwarzenbach, 2020 and Wieder & Ossimitz, 2015
The benefits of Power BI	The benefits provide justification for incorporating Power BI into the method.	Sinha, 2021 and Rayarao & Donikena, 2025
Data modeling	Data modeling principles provide guidance for data structure.	Sinha, 2021 and Ballard et al., 2006
Dashboard design principles	These principles offer a foundation for creating an informative and relatable dashboard.	Ivanković et al., 2021 and Bach et al. 2023
Lakehouse functionalities	These functionalities support the justification for using a lakehouse approach.	Armbrust et al., 2021; Schneider et al., 2024 and Sundar et al., 2022

Table 2. Components that require modifications for the method.

Component	Explanation	References
Warranty related data	Some mentioned warranty data can be used, but the dataset requires refinement for the current context.	Wallace et al., 2011; Wu, 2013; Marshall et al., 2018; Blischke, 1993 and Annadurai, 2023
WDA process	The WDA process provides a useful structure, but its stages require adaptation for this study's context.	Annadurai, 2023
Data quality definitions	The definitions provide criteria that can be applied within the scope of the study when specifying the method's data quality requirements.	Miller et al., 2024 and Wang & Strong, 1996
ETL and ELT	These processes are relevant, but their steps must be adjusted to match the requirements for the method.	El-Sappagh et al., 2011; Vassiliadis, 2009; Simitsis et al., 2023; Ponniah, 2010; Biswas et al., 2019; Khan et al., 2024; Storey & Song, 2017; Jo & Lee, 2019 and Dmitriyev et al., 2015
Chart selection	Certain chart selection guidelines can be applied.	Sherman, 2014; Saket et al., 2019; Sedrakyan et al., 2019 and Srinivasan et al., 2019
Chart layout	Layout principles will be applied where they align with the visual and analytical goals of the method.	Zhang et al., 2024

2.6 Research gap

As shown in the literature review, many individual aspects relevant to this study have been discussed in prior research. Existing studies address topics such as data quality, the variety of warranty-related data that can be analyzed, the functioning and implementation details of ETL and ELT processes, the use of lakehouse architectures with attention to data quality, and dashboard design practices. However, no studies were found that combine these aspects into a single method for ensuring warranty data quality and processing within a lakehouse environment while enabling dashboard visualization.

The only peer-reviewed study that comes close to addressing a similar setup to this study is from Sreedhar et al. (2024). In their research, they studied two approaches of processing data using Apache Spark in Databricks and visualizing it in Tableau, as well as analyzing the data using Azure services and creating visualizations in Power BI. Although this research demonstrates the use of multiple tools and introduces steps for the method, it falls short in key aspects relevant to this study. The study does not provide a deep methodological explanation regarding the data pipeline itself, its quality assurance practices, or the principles behind the visualization decisions. Additionally, the research does not utilize warranty data, meaning it does not address warranty-specific considerations and therefore cannot support warranty decision making.

However, there are studies that have directly addressed the visualization of warranty data. Annadurai (2023) examines how warranty data can be visualized by presenting multiple visualization types such as failure descriptions, months in service at failure, usage patterns, seasonal failure trends, and probability plots. The study also discusses how these visualizations reveal warranty insights. In a master's thesis, Patel (2025) developed a dashboard utilizing warranty data related to maintenance and faults. The resulting interactive dashboard combines multiple visualization types, including bar and line charts, scatter plots, and bubble diagrams. The dashboard enabled users to detect failure patterns, examine differences between categories, and monitor changes over time, and the

evaluation indicated that these visualizations were valuable for decision-making. Ojala (2024) studied also in a master's thesis how predictive modelling and AI integrations can be utilized for warranty and notification data analysis. The result introduced a Power BI dashboard in which the data was visualized and analyzed using predictive analytics. The dashboard included various visualizations such as tables, line graphs, key influencers and bar charts. However, none of these studies incorporate data quality as part of their development process.

Taken together, the existing research shows that individual components have been studied separately, but not as an integrated whole. What remains missing is a comprehensive method that combines data from different sources within a lakehouse architecture, incorporates data quality considerations, and enables data visualization. This gap highlights the need for a unified approach, which this study aims to address. In addition to the academic gap, there is also a practical need in the organization involved in this research. The warranty management requires a new Power BI dashboard that can efficiently present warranty-related insights and integrate climate information, enabling improved decision-making. In addition, the underlying process should be clearly defined and documented. By addressing both the research gap and the organization's needs, this study aims to contribute to the knowledge base and solve concrete issues.

3 Design science research

As the aim of this research is to develop a method for building a new analysis dashboard that is tightly connected to the business process and addresses organizational challenges, design science research (DSR) was chosen as the methodology. This choice aligns with Hevner et al. (2004, p. 77), who state that design science in the information systems (IS) field is concerned with solving identified organizational challenges by developing and assessing IT artifacts. Through this process, design science researchers gain insight into the problem targeted by the artifact and evaluate the proposed solution. Furthermore, Walls et al. (1992, p. 42) argue that design encompasses both the process, which is a group of actions, and the artifact, which represents the outcome and insight generated through the process. According to Hevner et al. (2004, p. 78), this perspective of design supports a problem-solving paradigm in which attention alternates between the design process and the resulting artifact within the same complicated problem. Therefore, it is essential for the design science researcher to understand that developing the process and the artifact are both part of the research.

Hevner et al. (2004, p. 77) explain that artifacts in DSR can range from programs and mathematical or formal logic representations to informal, natural-language descriptions, as long as they are presented in a structured form. They categorize these artifacts as "constructs, models, methods, or instantiations," which are widely accepted definitions of artifact types in IS design science research (Hevner et al. 2004, p. 82; Winter, 2008, p. 471). Constructs refer to the development of vocabularies and symbols that help articulate both the problem and potential solutions (Hevner et al. 2004, p. 82). They provide the language needed to describe and structure a problem area, which significantly influences design work (Hevner et al. 2004, p. 82; Winter, 2008, p. 471). Models use a set of constructs to represent problems and solutions (Winter, 2008, p. 471). Methods describe the processes that offer instruction for solving a certain problem. Instantiations demonstrate the capabilities of the design process and design product, meaning that instantiations are case specific combinations of constructs, models, and methods (Hevner et al. 2004, p. 82; Winter, 2008, p. 471). Hevner et al. (2004, p. 83) further note that artifacts

developed in design science are not usually fully developed and functional information systems. Instead, they are innovations that outline the ideas, practices, technical capabilities, and products that enable the analysis, planning, and implementation of information systems.

As this research focuses on explaining the process of developing a warranty data analysis dashboard, the artifact in this study is a method. By relating this approach to Winter's (2008, p. 471) definition of a method, the similarity becomes clear. The study outlines the steps involved in developing the warranty analysis dashboard (i.e. describing processes) to provide new and reliable insights into warranty data (i.e. solving a certain problem). Hevner et al. (2004, p. 87) explain that the development of new and appropriately evaluated methods that improve the existing knowledge base represents an important research contribution if the artifact depicts precisely the technical environments while addressing previously unsolved problems.

Hevner (2007, p. 88) argues that DSR requires iterating through and clearly presenting three research cycles: the relevance cycle, the rigor cycle, and the design cycle. The relevance cycle begins by defining research requirements, such as the problem to be solved and the acceptance criteria for the final evaluation. It identifies the application context and represents the opportunities and problems present in the environment. After the artifact is introduced into its environment and demonstrated, the results determine whether a new iteration of the relevance cycle is needed. The rigor cycle ensures that the research contributes new knowledge and maintains innovativeness (Hevner, 2007, p. 90). It prevents the research from becoming routine design by referencing and expanding the existing knowledge base (Hevner et al., 2004, p. 81). The new results added to the knowledge base may include extensions of existing theories and methods, invented design solutions or processes, or insights and experiences obtained while conducting the research and testing the artifact in its real context (Hevner, 2007, p. 90). In addition, Hevner et al. (2004, p. 88) state that rigor also requires grounding the study in theoretical foundations as well as empirical work. The design cycle, described by Hevner (2007, p.

90) as the core of DSR, involves rapid iterations from building the artifact to redesigning it based on evaluation and feedback. While being dependent on the relevance and rigor cycles, it also operates independently in the practical implementation of the research. The design cycle should be based on relevance and rigor, but it can also introduce new elements to them through its own iterative process.

3.1 Design science research methodology

Peffers et al. (2007) translated Hevner's design cycles into design science research methodology (DSRM), which provides six practical steps for producing and presenting DSR, as illustrated in Figure 6. As noted earlier, it is important for the design science researcher to understand that both the development of the artifact and the development of the process are part of the research (Hevner et al., 2004, p. 78). Therefore, this study aims to develop a process by following the steps of DSRM.

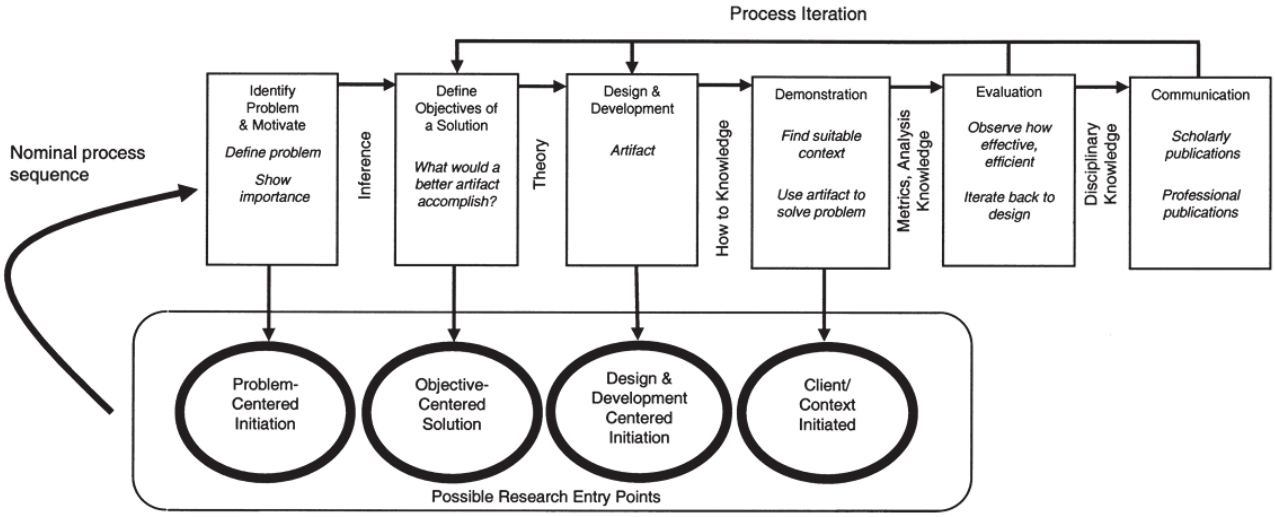

Figure 6. DSRM process model (Peffers et al., 2007, p. 54).

The first activity presented by Peffers et al. (2007, p. 52) is called "problem identification and motivation". In this activity, the research problem is clearly defined and its value is explained. Since the problem definition guides the development of an artifact intended to provide a solution, it is beneficial to divide the problem into smaller, logically

structured parts without neglecting critical details of the problem complexity. Addressing the value of solving the problem serves the purpose of motivating both the researcher and the intended audience to aim for the solution, accept the outcomes, and clarify the reasoning behind the researcher's interpretation of the problem. In this activity, the researcher should understand the current state of the problem and the significance of resolving it.

In this research, the first activity is addressed in the introduction, the research gap section of the literature review, and the initial situation section of the results. The research problem focuses on the need to develop a new process for generating new and reliable insights using appropriate technologies. The value of the proposed method lies in its ability to provide guidance for organizations, while the insights generated from the resulting dashboard support warranty-related decision-making and analysis. Furthermore, the identified research gap highlights the need for a method that integrates the selected components into a coherent process. In addition, as the company involved is undergoing a transformation of its data architecture, the proposed approach must align with this transition.

The second activity is "Define the objectives for a solution" (Peffers et al., 2007, p. 55). The objectives should be defined logically based on the results of the first activity and by considering what is possible and achievable. These objectives may be either quantitative or qualitative, depending on whether they specify measurable improvements over existing solutions or describe the artifact's capability to address problems that have not previously been solved. Furthermore, the current state of the problem and available solutions must also be taken into consideration in this activity.

The objectives of this research are defined by the research questions and the requirements section of the results chapter. The research questions indicate that the objectives involve identifying suitable components from prior research, outlining the steps of the method, and demonstrating its applicability for addressing the identified problem. The

requirements further specify the objectives that need to be achieved. By establishing a method, the process becomes repeatable, enabling the identification of both inconsistencies and effective practices.

The third activity “design and development” (Peffers et al., 2007, p. 55), focuses on creating the artifact. As previously noted, the artifact may be a construct, model, method, or instantiation. However, within DSR, the artifact may be any designed object that establishes a research contribution. This activity involves defining the functionalities and architecture of the artifact and developing the specified solution. Completing this activity requires drawing on knowledge abducted from prior research.

The design and development of the artifact are presented in the artifact design and development section of the results chapter. In this section, the method artifact is introduced, describing the process of the solution. The design decisions are grounded in prior research as well as insights from the experts in the organization. These inputs emerged from stakeholder interviews and internal requirements. As the design cycle is an iterative process, the prototype is continuously refined based on stakeholder feedback to ensure that it evolves in the desired direction.

The fourth activity is “demonstration” (Peffers et al., 2007, p. 55). In this stage, it must be illustrated how the artifact can be applied to solve one or more instances of the identified problem. The result of the artifact may be demonstrated through simulations, case studies, or other suitable activities. Proof-of-concept prototypes may be utilized at this stage, as demonstrated in the case examples presented by Peffers et al. (2007, pp. 59, 66). Demonstration requires a solid understanding of how to use the artifact to resolve the problem (Peffers et al., 2007, p. 55).

The demonstration is presented in the demonstration section of the results chapter. In the context of this research, it is carried out by illustrating how the proof-of-concept method and the resulting dashboard prototype address the identified problem. This is

achieved by comparing the requirements with the method and by having stakeholders compare the dashboard against these requirements. Such comparisons provide evidence of the method's practical applicability and its ability to meet the specified requirements.

The fifth and sixth activities defined by (Peffers et al., 2007, p. 56) are "evaluation" and "communication". Evaluation is conducted by measuring how successfully the artifact performs in solving the problem, thereby demonstrating its value. Based on the evaluation results, the researchers determine whether further refinements are necessary, which may require returning to the design and development activity, or whether improvements should be left to future research. The communication activity represents the final phase, in which the research is published for researchers and other relevant stakeholders. However, the last two activities fall mostly outside the scope of this study. This limitation is due to time constraints and the predetermined scope of the study. Only the communication phase will be completed on a smaller scale by presenting the research and its results to the company's warranty team, other interested stakeholders, and the university.

3.2 The research environment

The research was conducted in a real organizational setting within a large multinational company, specifically within its energy business operations. The energy business operates globally, with products delivered to customers in 180 countries. Customers can choose from various products and implementation options based on their requirements and needs. The scale and complexity of this environment provide a rich context for data-driven analysis.

The study focused on the warranty services team, particularly its management, which identified a need for a new data visualization dashboard and a structured process for its development. The warranty organization is responsible for all warranty-related activities. This includes monitoring and processing warranty claims, analyzing root causes and costs,

and coordinating corrective actions with both internal and external stakeholders. In practice, the team acts as the central point of contact for any issues arising during the contractual warranty period. In addition, the warranty organization is responsible for reclaiming and delivering feedback to support product improvements related to warranty issues. The team already utilizes systems for WDA to provide high-quality insights into performance of the organization's products. The systems continue to be developed to implement more advanced methods, provide deeper insights, unify system usage, and minimize the number of separate systems required.

3.3 Data collection

In this study, data collection was conducted to support the design, development, and demonstration of the proposed method in accordance with the DSR approach. First, existing academic literature and prior research were reviewed to identify relevant components suitable for the method. The literature search was conducted using multiple scientific databases, including ScienceDirect, Web Of Science, IEEE Xplore, and AIS eLibrary. In addition to academic sources, expert knowledge was gathered through informal discussions with the experts within the participating company. These experts possessed practical experience in warranty process, data analytics, and business intelligence. Their input contributed to shaping the method by providing insights into data availability, constraints, and practical requirements.

The empirical data used to develop and demonstrate the method consisted of two datasets. Warranty data was obtained from the company's internal data storage systems and files and included historical records related to warranty cases and associated supporting information. Climate data was collected from external data provider offering climate zone classification with corresponding geographical locations (Yu et al., 2023). Both datasets represented secondary data, as they were originally collected for purposes other than this research.

4 Results

This chapter explains how the resulting method was constructed and applied in practice. It begins by outlining the initial situation and the requirements that guide the design, as these define the problem context and criteria for the method. The chapter then details each step of the design and development process, including data selection, platform and ELT decisions, data modeling, data quality considerations, and visualization, to ensure the repeatability of the method. It also describes the iteration and refinement activities during development, highlighting how the feedback and testing influenced the final solution. Furthermore, the method is summarized in a table, which enhances reproducibility and provides a clear structure for its implementation. Finally, the chapter presents the demonstration of the proposed method to show its applicability in addressing the identified problem.

4.1 Initial situation

The company involved in this research is currently undergoing a data-driven transformation, with an increasing emphasis on improving data quality, unifying system usage, facilitating data sharing and usage among employees, and increasing contributions to artificial intelligence. Historically, the organization relied on a combination of enterprise data warehouse (EDW) as data storage and QlikView for reporting. Later, the company began adopting Power BI as both a data platform and a visualization tool, using it in combination with the existing EDW. However, maintaining and storing data directly within Power BI proved to be costly and inefficient in the long term. To address these limitations, the company is evaluating the role of Databricks, a modern data lakehouse environment, to determine its suitability in resolving existing data challenges. Consequently, the overall direction is shifting toward an architecture where Databricks is used for data storage and processing, while Power BI focuses on reporting and visualization.

The warranty team is keeping pace with these organizational changes and aims to enhance data-driven decision-making. While the team already utilizes some Power BI

dashboards, they have identified the need for a new data visualization dashboard. Specifically, there is a need to analyze the impact of varying climate conditions on products and integrate this information with warranty-related data. As the products are highly affected by the surrounding climate, such insights would help the warranty management identify correlations between climate factors and reasons for claims, analyze costs across different climates, justify why certain product components fail in specific environments, and, in the long run, contribute to improving product quality.

4.2 Requirements

The requirements for the artifact were defined to support the development of a structured and repeatable method. The requirements were derived from the needs expressed by the warranty management and complemented with requirements identified in relevant literature. While the organizational requirements ensured that the method addresses the practical needs and therefore contributes to the relevance cycle of DSR, the requirements derived from literature grounds the work in existing scientific knowledge, thereby supporting the rigor cycle of DSR.

From a warranty management perspective, the method must ensure data quality, meaning that the underlying data can be fully trusted. Data quality is a prerequisite for drawing valid conclusions and making correct decisions. To align with the organization's strategic direction, Databricks should be utilized to perform the required data processing workflow. The data visualization stage should be executed in a Power BI environment, with the aim of developing a dashboard that allows users to interact with the data through filtering and by selecting specific subsets of information. Warranty management has defined a set of insights they aim to obtain from the final dashboard, which guides the selection of the data. The dashboard should provide a clear interface that supports efficient data exploration and interpretation. At this stage, automated data refresh is not required, as the method is a proof-of-concept implementation. This type of execution is supported in design science as mentioned by Hevner et al. (2004) that artifacts are not

usually fully operational systems but rather innovative constructs that enable analysis, planning, and implementation of information systems.

The literature review identified several requirements that guide this study, particularly regarding warranty data, data quality, ELT processing, dashboard design, and the data lakehouse. Requirements for warranty data specify that it should be collected and processed with high quality, as incomplete and imprecise data can have a negative effect on data-driven decision making. More broadly, data quality requirements underline the need to consider multiple attributes referred to as usefulness, quantity, semantics, intrinsic quality, and representational quality that determine the overall quality of the data. In this way, the true value of the data can be effectively utilized. Data quality requirements should also be supported by the lakehouse's medallion architecture, which enables dividing the data into layered structures based on the level of refinement. ELT theory provides requirements on how data should be extracted, loaded, and transformed. ELT should be accomplished in the lakehouse environment to accelerate onboarding of new data sources and reduce the complexity of data pipelines. Dashboard implementation requirements guide the selection of appropriate visualization types, layout decisions, design principles, and design trade-offs.

4.3 Artifact design and development

This section discusses the design and development of the method artifact. This activity focuses on defining the artifact's functionalities and developing the proposed solution in accordance with the DSRM guidelines outlined by Peffers et al. (2007). The artifact is constructed by combining insights obtained from domain experts within the organization with relevant components and principles identified in prior research.

4.3.1 Data selection and data sources

Prior research identifies data extraction as the initial phase of ETL/ELT and WDA processes (El-Sappagh et al., 2011; Vassiliadis, 2009; Simitsis et al., 2023; Annadurai, 2023).

However, this phase requires defining the data to be retrieved. Consequently, the selection of relevant data attributes was conducted as an essential initial step. This selection was driven by the stakeholders, who defined the insights and analytical views they wish to obtain from the analysis. Based on these requirements, the data attributes were selected by identifying the data needed to enable the requested insights. Several of these attributes are supported by prior research, which also provides established classifications for organizing the selected data.

The literature identifies a high-level classification of warranty data into warranty claim data and supplementary data, which provides structure for organizing the data (Wallace et al., 2011). The insights related to warranty claim data identified by the stakeholders include amount of claims and amount of warranty costs, analyzed across different warranty types, areas, and years. These insights are supported by prior literature, which specifies variables commonly used to analyze warranty claims, such as the number of products with claims occurring within specific time frames, the date of the warranty claim, frequency of claims, as well as total warranty costs broken down by year, region, and customer (Wu, 2013; Marshall et al., 2018; Blischke, 1993; Annadurai, 2023). Supplementary data supporting these insights, as identified by the stakeholders, include climate zone, project type, product type, failure component group, classification, coordinates of the location, country, and location name. The selected data attributes form the basis for the final analysis and define the insights that are presented in the dashboard.

The next step consisted of identifying the source systems where the required data was located. The source systems were identified with the assistance of data analysts familiar with the company's data infrastructure and the availability of external climate data. The majority of the data was stored in the company's enterprise data warehouse (EDW), where fact tables and dimensional tables had already been defined. Climate zone data was obtained from an external Python package that enables the identification of climate zone for any given location. Latitude and longitude data, indicating the location of the product, was available within the Databricks environment. In addition, an Excel file

containing static reference data was identified to determine the specific location of failures within individual components of the product. Table 3 summarizes the identified data sources, their corresponding dataset names, and a brief description of their contents to support understanding of how data selection can be systematically conducted within this method.

Table 3. Overview of data sources, datasets, and their descriptions.

Data source	Dataset name	Description
EDW	Notifications	Warranty claims and related event data
EDW	Installation	Product data
EDW	Equipment	Data about specific components in products
EDW	Country	Specifications of areas and regions
EDW	Project	Project classification data
EDW	Volumes under warranty	Data specifying products under warranty
EDW	Calendar	Calendar data to enable time-based analysis
Databricks	Installation location	Latitude and longitude of product locations
Internal Excel file	DLC groups	Location codes of failures within individual components of the product
External data source	Climate zone Python package	Climate zone classification based on geographic location

4.3.2 ELT platform selection and implementation approach

When the relevant data and their sources had been identified, the next step was to implement the ETL/ELT process in a suitable environment. As data originated from multiple sources, a unified platform was required to combine, store, and implement transformations. In this method, these tasks were carried out in a Databricks lakehouse

environment, which had been defined by the stakeholders as the required platform for data processing. The choice of Databricks was further supported by findings from the literature review.

The literature highlights that the lakehouse architecture provides support for business intelligence workloads, which aligns directly with the objectives of this study (Armbrust et al., 2021). In addition, prior research indicates that traditional architectures combining separate data lake and data warehouse solutions often lead to increased cost and complexity, whereas the lakehouse model addresses many of these issues by providing a unified platform for both data storage and analytics (Simitsis et al., 2023; Armbrust et al., 2021; Schneider et al., 2024). However, as mentioned, EDW is utilized as one of the sources for data which is in contrast with this finding. This design choice can be explained by the fact that EDW contained data that was already in a structured and mostly clean format, making it reasonable to utilize. Consequently, the use of both systems reflects the current transitional state of the organization's data architecture rather than an idealized target design.

Once the lakehouse environment had been selected as the processing platform, the focus shifted to determining the most appropriate data integration and transformation approach within this architecture. As noted by Sundar et al. (2022), although traditional ETL remains applicable in the lakehouse environment, ELT has increasingly become the preferred approach. Their research indicates that modern data teams often adopt a hybrid model of implementing minimal ETL, such as the handling of personally identifiable information, at the edge, while core logic and modelling are performed through ELT. This reduces pipeline complexity and accelerates the integration of new data sources.

Furthermore, Dmitriyev et al. (2015) note that ELT is well suited for business intelligence due to its flexibility and adaptability, while Simitsis et al. (2023) highlight its suitability for environments in which data volumes grow continuously and are often generated without direct human intervention. Cloud platforms, such as Databricks, strengthen this

trend by enabling cost-efficient processing of distributed data and by decreasing computing costs. In addition, ELT provides flexibility for integrating new data sources, since transformations occur inside the target system which also allows raw data to be reprocessed multiple times if required (Dmitriyev et al., 2015). For these reasons, ELT was selected for use in this study, as it provides flexibility, cost-efficiency, and support for data integration and processing needs of the method.

4.3.3 Extract and load

In the extract and load phase, the prior research highlights that the characteristics of each data source must be considered to ensure adequate extraction (El-Sappagh et al., 2011). The literature on data loading describes three commonly used approaches referred to as initial, incremental, and full refresh (Khan et al., 2024). Simitsis et al. (2023) further note that, within ELT architectures, data load and extraction often function primarily as replication mechanisms. These findings are addressed in a manner appropriate for a proof-of-concept method.

In the context of this study, EDW is updated on a daily basis, making incremental loading to Databricks the preferred approach, as it imports only the changes in data at scheduled intervals. However, during this research, the integration between the company's EDW and Databricks was not fully operational, which required the extract and load phases to be executed as a data replication using non-refreshing data. This limitation was considered acceptable for the purposes of this study, as the implementation served as a proof-of-concept. Similarly, Python package and the Excel file used in the solution are not expected to change in the near future. Therefore, they were ingested to Databricks using a one-time initial load, without the need for scheduled refresh operations given their current stability. Furthermore, data already available within the Databricks environment, maintained and updated by other teams within the organization, was utilized directly without additional extraction or loading steps.

4.3.4 Transform

After extracting and loading the data into Databricks, transformations were executed within the platform to leverage its computational capabilities, fulfilling stakeholder requirements and aligning with ELT principles. Performing transformations ensures that data becomes accurate, consistent, and precise (Khan et al., 2024). The transformation phase incorporated several commonly used transformation techniques identified in the literature, including cleaning data, dataset joining, conversion, summarization, and the application of business logic (Ponniah, 2010; El-Sappagh et al., 2011; Sundar et al., 2022). The transformed data was stored in separate tables to maintain the availability of original datasets for future use and organizational collaboration, consistent with Databricks' collaborative environment (Gaur et al., 2025). During data transformation, the natural language assistance in Databricks was used as a supporting tool for query formulation.

Data cleaning activities were performed during the transformation phase, as these steps were not included in the extract and load stages in accordance with the ELT approach. Although the source data from EDW was generally available in a structured and well-maintained format, previously identified data quality issues necessitated additional cleaning efforts. As a result, certain rows were removed from the dataset to eliminate irrelevant records. Additionally, conversions were applied to primary and foreign keys to ensure consistency in data types and to make them suitable for use in the dimensional model.

Join operations were applied mostly for datasets that did not independently form complete analytical entities. These joins were implemented between geographical locations and climate zones and between notifications table and volumes under warranty. The join operations were designed to restrict the dataset to analytically relevant records or to enrich the data with contextual information essential for deriving accurate analytical insights. In addition, joins between the country and installation dimension tables were performed to preserve a star schema structure, supporting the alignment with dimensional modeling principles.

Summarization was applied to aggregate detailed data into a higher level of granularity. This included consolidating the detailed climate zone classifications into broader groups, enabling clearer filtering and interpretation in the analysis. In addition, detailed claim identification codes were summarized to a new column by extracting only the prefix, resulting in a simplified representation that supports analytical grouping.

The final transformation step involved the application of business logic. Within the method, this refers to ensuring analytical relevance and focusing the dataset on records suitable for the intended analysis. This design decision was necessary because the source data contained unnecessary business lines, claim types, and project types. Business rules were applied to retain only warranty claims, energy business related values, and selected project types. In addition, filtering was also performed for all tables to retain only the necessary columns, thereby reducing unnecessary computational overhead.

4.3.5 Visualization platform selection and data integration

As discussed earlier, the computational processing was implemented in Databricks, while visualization and reporting is required to be implemented through a separate platform. Power BI was selected for this purpose, as its use was defined as a requirement by stakeholders and it provides suitable capabilities to support the objectives of the solution. As mentioned, Power BI offers visualization functionalities through Power View, enabling the development of tables, charts, and maps required for analytical reporting in this study (Sinha, 2021). In addition, the drag-and-drop interface facilitates rapid construction of visual elements, supporting iterative refinement of the dashboard. Power BI also supports automatic data refresh, which is suitable given that the underlying data is expected to be updated later.

Furthermore, Power BI can establish a direct connection to Databricks compute clusters, enabling access to processed data (Rayarao & Donikena, 2025). This integration supports the overall architecture of the method by enabling a clear separation between data

processing and visualization layers. Data can be made available to Power BI either by importing the data or by querying it directly from Databricks. Domain experts recommended the use of direct querying, particularly for large datasets, which is also applicable in the context of this study. However, for the purposes of this proof-of-concept implementation, the import mode was selected, as it enabled faster query performance during the report development while also ensuring that Databricks query costs remained controlled during the experimental phase. For further development, direct querying is recommended to ensure efficient handling of increasing data volumes.

4.3.6 Dimensional modeling

As noted in prior research, a data model plays a central role in BI by defining and maintaining the structure of data, thereby enabling its effective use (Ballard et al., 2006). In this method, dimensional data modeling principles were applied using a star schema, selected for its better performance and understandability compared to the snowflake schema, as recommended by Ballard et al. The data processed and transformed in Databricks was structured according to a fact- and dimension-based table design and final data modeling was implemented on the Power BI side. This included defining the primary and foreign keys and establishing the relationships between tables in accordance with the star schema. Figure 7 presents the implemented data model and the relationships between tables, where the star schema is clearly visible, serving as an example of how such structure can be constructed.

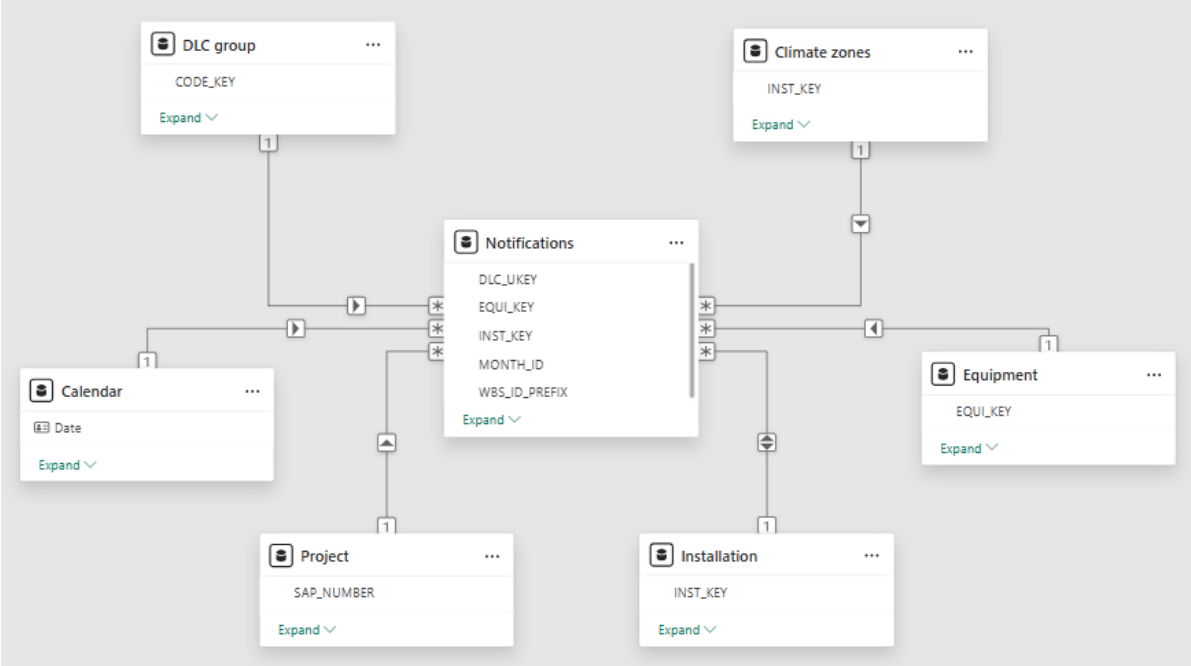

Figure 7. Data model implemented in Power BI.

4.3.7 Dashboard design

The literature review revealed that one of the objectives of business intelligence is to present high-quality information in a form that effectively supports decision-making (Wieder & Ossimitz, 2015). Prior research suggests that this can be achieved by selecting appropriate chart types, choosing suitable layouts, balancing key parameters, and applying design principles. In this method, Power BI was used for the visualization component, where theoretical insights were applied during the dashboard development process. The aim was to present the required insights through the dashboard while ensuring clarity and ease of interpretation.

As identified in the literature review, the selection of visualization types should be guided by the type of analysis being performed, an approach that was followed in the development of this method (Sherman, 2014). Warranty costs and claim amounts were chosen to be presented using stacked bar charts. This aligns with comparative analysis, where the heights of the bars enable direct comparison (Sherman, 2014; Saket et al., 2019). At the same time, the bars illustrate how different segments (DLC group 1, DLC group 2,

OTHER) contribute to the total values, thereby supporting contribution analysis, for which bar charts are suitable (Sherman, 2014). Warranty costs and claims were also examined across multiple years, which corresponds to time-series analysis, where bar charts are typically used to illustrate changes over time (Sherman, 2014; Sedrakyan et al., 2019). Location-based data is presented using map visualizations, as recommended by Sherman (2014) and Sedrakyan et al. (2019). In accordance with this approach, claim amounts were visualized on a map to demonstrate how the amounts of claims are distributed across regions. Climate zones were shown using a static map adapted from Kottek et al. (2006, p. 261), where the climate zones are visually illustrated. In addition, some tables were selected for the dashboard to support the information presented in the charts.

Design principles for dashboard creation introduced by Ivanković et al. (2021) further shaped the development process. Their study underscores the need to clearly identify the target audience and to limit the number of indicators to emphasize the importance of information. They also recommended presenting information from general to specific or arranging it by thematic categories to establish an intuitive flow of information. Additionally, the study highlights the necessity of maintaining transparency in data sources and methodologies. Furthermore, Bach et al. (2023) discussed in their research that page layout requires careful planning, including determining the number of pages, the content each page contains, and how the components are positioned. Zhang et al. (2024) found that the optimal dashboard layout places the primary chart in the left-center area, with the supporting non-core charts arranged in a partially symmetrical pattern around it. Furthermore, interactive features can be utilized to keep the content concise, allowing users the ability to reveal additional information only when they choose to explore it (Bach et al., 2023). This can be implemented through navigation buttons, tabs, links, or detail-on-demand features, where data remains hidden until triggered by user interaction.

The development of the warranty dashboard was further guided by these established design principles. The design process continued by defining the target audience as warranty management, which guided the selection of indicators. To maintain clarity, the dashboard included only essential indicators relevant to warranty management. The dashboard layout was structured so that the primary chart, warranty costs, was positioned in the left center area, with the second key chart, claim amount, placed next to it. Supporting visualizations were arranged around these charts in a partially symmetrical manner. This layout also supports progression from general to more specific information, moving from overall cost trends to detailed geographical and categorical breakdowns.

Interactive features were incorporated to present additional information on demand, enabling users to access further details only when required. These features include information buttons that provide explanations of climate zone classifications and calculation rules, as well as interactive filtering options that allow users to adjust the scope of the analysis. The filters included climate zone, project type, product type, DLC group, project classification, area, and country, as specified by the stakeholders to provide supplementary data. In addition, the display of the last data refresh timestamp and the availability of explanatory information supports transparency and builds user trust in the presented data. Figure 8 presents the design of the dashboard, where the design principles are implemented.

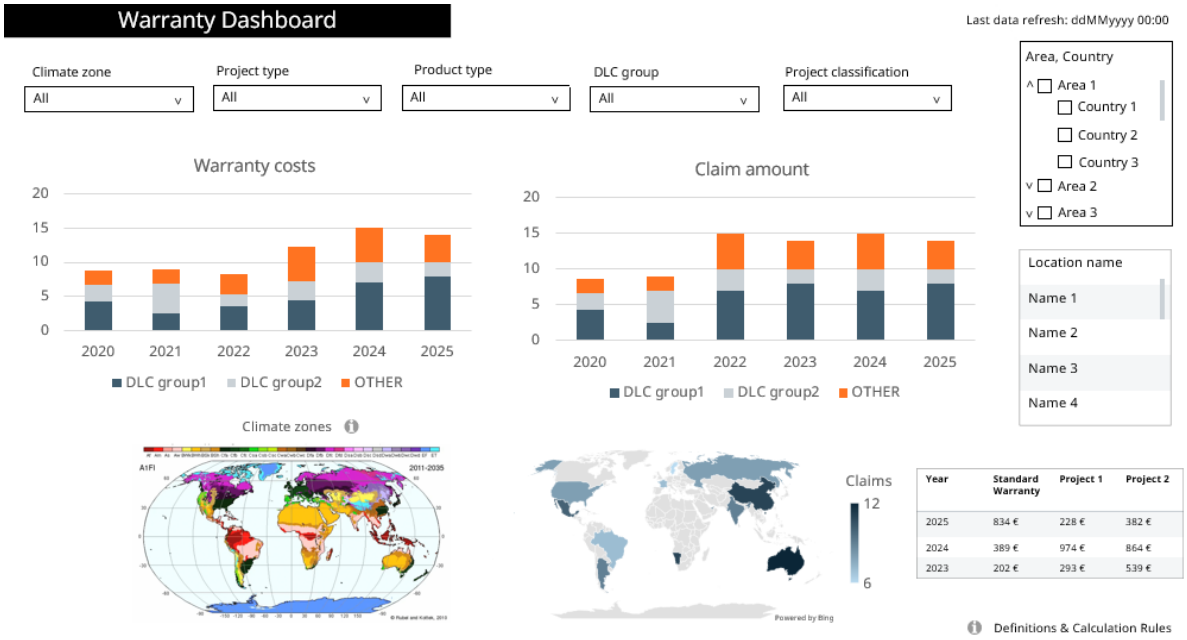

Figure 8. Structure and key components of the proposed dashboard.

4.3.8 Data quality

Earlier research places strong emphasis on the importance of data quality. While data is recognized as a central resource for organizations, its usefulness is not only based on the volume (Mahanti, 2019; King & Schwarzenbach, 2020). Rather, it depends on how organizations manage and ensure the quality of data. When data quality is inadequate, the effectiveness of software tools cannot compensate for the limitations created by poor-quality inputs. Therefore, it is necessary to inspect how data quality is considered in this method, ensuring that the artifact can operate on reliable and well-structured data.

As discussed, Databricks applies a medallion architecture consisting of bronze, silver, and gold layers (Databricks, 2025). This structure can be used to assess and manage data quality throughout the data lifecycle. In the context of this study, the only dataset that belonged in the bronze layer was the Python package as it represents external raw data that does not follow the organization's internal schemas and includes unvalidated CSV and PNG files. Once this external data was combined with data from the EDW, it could be promoted to the silver layer, as it then aligned with the structured format of the EDW

data. However, the bronze layer was preserved to ensure that the raw data remains accessible for future inspection or for potential alternative use cases, especially since Databricks enables shared data access across an organization (Gaur et al., 2025).

Although EDW data has already undergone most of the necessary cleansing within the organization's EDW processes, it was integrated into the silver layer after performing the mentioned minor additional data cleanings. The examined Excel dataset was already structured into rows and columns with clean values, which aligns with the criteria of the silver layer and could therefore be placed there. As described, the gold layer contains enriched datasets with aggregations (Databricks, 2025). Therefore, when business rules were applied, the data appropriately belonged to the gold layer. As a result, the data was fully prepared for business intelligence purposes, and the computational capabilities of Databricks had been leveraged to their fullest potential.

The method also addresses data quality aspects derived from the frameworks of Miller et al. (2024) and Wang and Strong (1996), focusing on the attributes of usefulness, quantity, semantics, intrinsic quality, and representational quality. Usefulness was achieved by implementing data transformations that addressed the defined user requirements and by curating the data specifically for visualization in the dashboard. The adaptability and reusability aspects of usefulness were supported by organizing the data according to the medallion architecture, which preserves both raw data and data enriched through transformations and business logic, thereby enabling future use in alternative analytical contexts.

The medallion architecture also supported representational data quality by organizing the data into clearly defined layers, enhancing interpretability and enabling meaningful results. The quantity attribute was addressed by selecting the required range of years and including all relevant datasets displayed in the dashboard, ensuring sufficient coverage for detailed analysis. Semantic quality was achieved by defining table and column names in an informative and consistent manner, supporting a clear understanding of the

data. Finally, intrinsic data quality was addressed by ensuring the correctness and credibility of the data through the use of validated data sources, integrating data with other reliable sources, and performing limited data cleaning based on predefined validation rules.

4.3.9 Iteration and refinement

As the DSR design cycle described by Hevner (2007) involves rapid iterations from building the artifact to redesigning it based on evaluation and feedback, the components of the method were iterated and refined continuously. The first component to be iterated was the data selection. This was due to the fact that the later availability of visual representations facilitated clearer discussions with stakeholders and made it easier to identify which data elements were relevant for reporting purposes. Consequently, the data selection was refined to correctly reflect stakeholder needs. This refinement also required modifications to the underlying tables to ensure that the appropriate data elements were included.

In addition, the data transformation activities were revised several times during development. Initial transformations were implemented and tested primarily to assess feasibility. These transformations were subsequently refined to improve correctness, performance, and alignment with the analytical requirements of the artifact. Iterative refinements were also applied to the dashboards based on feedback from stakeholders and domain experts. User feedback revealed both visualization related improvements and issues related to data correctness. When unexpected results were identified, the causes were investigated and corresponding adjustments were made to the data transformations and data selection logic. Furthermore, visual elements and filtering options were modified iteratively to support user needs and improve overall usability.

4.4 Method artifact overview

The method is illustrated in Table 4, which enhances reproducibility and provides a clear approach to its implementation. The table presents a brief description of each phase, the inputs required for execution, and the outputs produced in each phase. Data quality phases are integrated throughout the table.

Table 4. Overview of the method artifact.

Phase	Description	Input	Output
Data selection and source identification	Selecting data based on required insights and identifying internal and external data sources.	Stakeholders: Required data specifications and knowledge of data sources Literature: Warranty data, division to claim and supplementary data	List of data attributes and identified data sources
Data processing platform selection	Choosing Databricks lakehouse as the data processing platform.	Stakeholders: Requirement to use a lakehouse Literature: Benefits of using a lakehouse	Documented rationale for utilizing a lakehouse
ELT approach	Selecting and justifying the ELT approach.	Literature: Recommendations for choosing ELT	Documented rationale for adopting ELT approach
Extract and load	Extracting data from sources and loading it into the Databricks lakehouse without support for data refreshing.	Stakeholders: Limitations from the proof-of-concept implementation Literature: Approaches for extraction and loading	Data ingested into Databricks lakehouse to bronze and silver layers

Phase	Description	Input	Output
Transform	Transforming data through cleaning, joining, conversion, summarization, and application of business logic.	Stakeholders: Identified data quality issues Literature: Theory on data transformation types	Curated silver and gold tables
Visualization platform selection and data integration	Selecting Power BI as the visualization platform and integrating data from Databricks to Power BI using import mode.	Stakeholders: Requirement to use Power BI and integration recommendations Literature: Capabilities of Power BI	Documented rationale for utilizing Power BI and imported data ready for dashboard development
Dimensional modeling	Building the dimensional model in Power BI using a star schema.	Literature: Dimensional modeling theory and recommendations for star schema design	Data model
Dashboard design	Designing the dashboard by considering chart selection, layout, and design principles.	Stakeholders: Required insights Literature: Chart selection, design principles, and optimal layout	Dashboard prototype
Iteration and refinement	Iteratively refining data selection, tables, transformation activities, and the dashboard.	Stakeholders: Feedback on components of the method	Refined method that meets requirements

4.5 Demonstration

The method was demonstrated by comparing the proposed method and the resulting dashboard against the defined requirements. The final dashboard was developed in accordance with the method's defined steps, but it is not presented in this study due to the classified nature of the information it contains. However, it visually aligns with the

structure and functionalities illustrated in Figure 7, which provides an overview of the implemented solution. Warranty management compared the resulting dashboard prototype against the requirements to confirm that the method can be used to address the need for new information with reliable data quality. This aligns with the purpose of demonstration in DSR, which is to show how an artifact can be used to solve one or more instances of an identified problem (Peffers et al., 2007). In addition, the method was examined against requirements identified in the literature to demonstrate consistency with established research.

The requirement to utilize Databricks for data processing and Power BI for visualization was fulfilled by separating computational logic from the presentation layer. Databricks was used in ELT and preparing warranty and climate data, while Power BI was used to deliver insights through interactive visualizations. During the implementation, data modeling was closely connected to the reporting layer, which slightly blurred the separation between computation and presentation layers. This design choice did not prevent the proof-of-concept dashboard from being implemented as intended, but it highlights an area that could be further studied in future iterations of the method.

The ELT requirements were fulfilled by following the established ELT principles within the lakehouse. The extract and load phases were primarily implemented through data replication using non-refreshing data, while accounting for source-specific extraction and loading approaches to a level suitable for a proof-of-concept implementation. The necessary transformations, including data cleaning, dataset joining, conversion, summarization, and the application of business logic were performed within the lakehouse in accordance with established research on data transformation practices. The ELT approach enabled rapid onboarding of new data sources and reduced the complexity of the data pipeline.

The data quality requirements were addressed through the attributes of usefulness, quantity, semantics, intrinsic quality, and representational quality. Usefulness was

achieved through data transformations and the application of the medallion architecture, while representational quality was supported by the same architectural approach. Quantity was ensured by incorporating relevant datasets, semantics was enhanced through clear column definitions and table names, and intrinsic quality was improved through data cleaning and the use of validated tables. Only these attributes were followed due to the scope of the study. Therefore, warranty data and climate data were mostly collected and processed at a high-quality level, fulfilling one of the requirements, although further improvements are needed to address the remaining quality attributes.

The dashboard implementation followed requirements derived from the literature, with minor adaptations to ensure that desired insights and components were included. Appropriate chart types were selected based on the type of analysis and at a scale suitable for this implementation, meaning that not all chart types were included. Established design principles were applied to maintain clarity, ensure an appropriate layout, enable interactivity, and support transparency. Additionally, some components not explicitly required by the literature were included, such as a data refresh timestamp to enhance trust in the data and supplementary tables to support the information presented.

As part of the demonstration phase, warranty management tested the proof-of-concept dashboard to determine whether it fulfilled the defined requirements. Warranty management was asked a set of questions focusing on the dashboard's alignment with expectations, trustworthiness, included insights, support for decision-making, clarity, and potential development areas. According to the responses, the data presented in the dashboard was perceived as reliable, largely due to the transparency provided by the integrated definitions and calculation rules and the data refresh timestamp. However, some blank values were observed as a result of missing input data. This issue originates from the initial data collection phase, during which the method had not yet been applied, rather than from calculation errors. However, this did not weaken the overall analysis, as the missing values were limited in scope and did not significantly affect the key values.

From a content perspective, the dashboard included all required insights and components for the new warranty analysis, with a focus on warranty costs, claims, climate zones, and relevant filters. The dashboard was found to provide new information and to support decision-making by enabling the combination of multiple filters and the comparison of resulting insights across different conditions. The clarity of the dashboard was considered high, supporting effective understanding of the presented information. Although the dashboard fulfilled the specified requirements, the stakeholders also identified opportunities for further development in future iterations. For instance, the comparison of climate zones in relation to different filters can be improved for greater clarity. In addition, the inclusion of trendlines and more advanced filtering options would enhance the functionality. These enhancements would move the dashboard closer to functioning as a comprehensive analysis tool from a high-level perspective. Furthermore, the automatic refresh of external and internal data sources should be considered in future iterations, which would necessitate careful consideration of relevant data governance aspects.

Overall, by addressing the majority of the defined requirements, the demonstration indicates that the proposed method is applicable for developing this type of analytical dashboard. However, as some requirements were only partially fulfilled, further refinement is needed in future iterations. Table 5 summarizes the demonstration results by listing the requirements, how they were addressed, and the extent to which each requirement was fulfilled, thereby enhancing the clarity and transparency of the demonstration outcomes.

Table 5. Summary of the demonstration.

Requirement	**How requirement was addressed?**	**Requirement met?**
Utilizing Databricks for data processing and Power BI for visualization	The requirement was achieved by separating computational logic from the	Partially

Requirement	How requirement was addressed?	Requirement met?
	presentation layer. However, the distribution of data modelling between the platforms could be further studied.	
Following ELT principles in the lakehouse environment	ELT principles were followed at a level suitable for a proof-of-concept implementation within the lakehouse, which accelerated the onboarding of new data sources and reduced pipeline complexity.	Yes, at a level suitable for a proof-of-concept implementation
Following dashboard implementation practices	Established principles were followed regarding the chart types, design principles, and optimal layout, to ensure clarity and information understanding. Minor adaptations were made to support the information presented.	Yes, with minor adaptations
Inclusion of required insights and components	The dashboard included all the required insight and components related to warranty costs, claims, climate zones and filters.	Yes
Data quality	Data quality was addressed through several defined quality attributes and the use of the medallion architecture, although additional attributes could be incorporated.	Partially

5 Concluding remarks

This research addressed the problem of developing a method for analyzing warranty data in combination with climate data. The objective of the study was to design a method artifact that integrates internal warranty data with external climate data using a lakehouse architecture, while ensuring data quality and supporting warranty management decision-making through an interactive dashboard. The study followed the DSR methodology, by applying the process model proposed by Peffers et al. (2007). Conducting the research within a real organizational context supported the relevance cycle of the developed artifact. The rigor cycle was ensured by grounding the method development in prior research and extending the knowledge base through the proposed process. Furthermore, the design cycle was supported through the iterative development of the method while being based on relevance and rigor.

5.1 Addressing the research questions

The research questions were addressed throughout the study. The first research question, “What prior knowledge is available to inform the design of the method artifact?” was addressed through the literature review. In addition, Tables 1 and 2 summarized the identified components that could be directly utilized as well as those requiring modification. The prior knowledge included best practices and established approaches related to warranty data analysis, data quality management, ELT and ETL processes, business intelligence, and data storage. These findings provided a foundation for the development of the proposed method.

The second research question was “What steps should the method include based on prior knowledge and input from company domain experts?”. This question was addressed in the artifact design and development. The method steps were derived from the components identified in prior research and complemented by knowledge from the organization’s experts. Overall, the method’s steps encompass data and source selection, data integration into a lakehouse environment using an ELT approach, data

transformation and data quality management, dimensional modeling, and visualization using Power BI. The method is summarized in Table 4 to ensure its generalizability and reproducibility.

The final research question, “To what extent is the developed method applicable for addressing the identified problem in practice?”, was answered in the demonstration section. The demonstration included comparing the defined requirements with the method and the resulting dashboard, to identify the method’s suitability for providing trustworthy information to support warranty decision-making. The results of the demonstration showed that the method mostly fulfilled the specified requirements and enabled the development of a dashboard that provides the desired insights. Stakeholder feedback further confirmed the practical applicability of the method in warranty management decision-making.

The results indicate that the individual components of the method successfully fulfilled their intended roles and collectively contributed to the solution. The ELT process implemented in the lakehouse environment proved suitable for many reasons. Performing ELT within Databricks reduced pipeline complexity and enabled efficient onboarding of new data sources, including external data such as climate data. In addition, the lakehouse architecture facilitated utilization of shared organizational data. Furthermore, defining data quality management within the lakehouse using the medallion architecture contributed to the trustworthiness of the analyzed data. The transfer of curated data to Power BI enabled efficient dimensional modeling and visualization. The final dashboard generated insights by combining climate and warranty data, demonstrating that the method supports warranty management decision-making and that the data presented can be considered trustworthy.

5.2 Main contributions and reflection on the research gap

The main contribution of this research is the development of a unified and practical method artifact for warranty data analysis. The contribution is grounded in a clearly

documented process that can be adapted to similar analytical contexts. Although the method was applied using warranty and climate data in this study, it is capable of being transferable to other analytical projects that require the integration of multiple data sources, data quality management, and dashboard development. This method is particularly relevant for organizations transitioning toward a lakehouse architecture and are aiming to separate data processing from the visualization layer. Within the context of warranty management, the method also contributes by uncovering reliable insights that may otherwise remain hidden.

The results of this study are consistent with prior research on warranty data analysis, data quality management, and business intelligence. Earlier studies emphasize the need for reliable warranty data, structured data quality process, and appropriate visualization principles to support decision-making (Wu, 2013; King & Schwarzenbach, 2020; Wieder & Ossimitz, 2015; Ivanković et al., 2021). The developed method reflects these principles by integrating warranty and climate data within a lakehouse environment, including data quality management into the analytical process, and following established dashboard design principles. In addition, the efficiency of ELT approaches and lakehouse architectures identified in prior research was also observed in this study (Armbrust et al., 2021; Sundar et al., 2022). However, instead of examining these components in isolation, this study demonstrated how they can be combined into a unified method within a real organizational context, contributing to addressing a gap in the existing research.

5.3 Limitations and future research

Even if the demonstration yielded positive results, this study has some limitations. First, the organizational context may limit the generalizability of the proposed method, as it was developed and demonstrated within a single organization using specific datasets. For instance, in this case the data was already relatively well structured, which simplified data validation and transformation tasks. Consequently, applying the method in organizations with less mature or more fragmented data environments may require additional effort. However, many established design principles were followed, which are not tied to

organization-specific implementations, thereby improving the generalizability of the proposed method. Second, the proof-of-concept prototype did not include automated data refreshing, which means that the solution was not yet fully operational. Introducing continuous data refreshing may reveal additional technical or data quality challenges that were not addressed in this study. Third, not all aspects of data quality could be thoroughly considered due to the scope of the research. For example, data governance issues related to external data sources were not fully evaluated. Given the sensitive nature of data quality, the method would benefit from a broader focus on data quality dimensions.

The concrete future development recommendations include further dashboard development, the implementation of automatic data refresh, improvements on data quality, and architectural considerations. Based on the demonstration phase, several opportunities to further develop the dashboard into a more comprehensive analytical tool were identified, such as enhanced filtering options, the inclusion of trendlines, and clearer climate zone comparisons. Automatic data refresh was also highlighted as a potential improvement, as it would introduce significant data governance considerations. Furthermore, the presence of blank values originating from the initial data collection phase, where the method was not yet applied, emphasizes that data quality must be considered throughout every phase of data usage. In addition, the method utilized a combination of a data warehouse and a lakehouse, reflecting a transitional architectural state of the organization that is not fully aligned with approaches recommended in prior research and should therefore be considered in future developments

In terms of future research, data quality should also be examined in greater depth and across a wider range of data types and complexity levels, such as with unstructured data. The proof-of-concept method should also be extended toward demonstrating its value by implementing the evaluation phase of the DSRM. In addition, future studies could further explore ELT and ETL practices within lakehouse architectures to identify best practices. The AI capabilities available in a lakehouse also present promising opportunities for extending the method. Furthermore, as noted in the demonstration phase, the

separation of computational logic and visualization could be enhanced by further examining dimensional modelling directly within the lakehouse. Overall, lakehouse technology is still emerging, and the availability of established research and best practices is limited. This contrasts, for instance, with BI implementations, where extensive prior research on their development already exists. Therefore, lakehouse technology represents a valuable topic for further research.

References

Al Olimat, S., Abu-Oliem, M., & Alkshali, S. (2025). Achieving Organizational Flexibility Through Business Intelligence at Jordan Customs. *Journal of Intelligence Studies in Business*, *14*, 62–76. https://doi.org/10.37380/jisib.v14.i2.2525

Annadurai, N. (2023). A Robust Warranty Data Analysis Method Using Data Science Techniques. *2023 Annual Reliability and Maintainability Symposium (RAMS)*, 1–6. https://doi.org/10.1109/RAMS51473.2023.10088226

Araujo, E. B. (2023). Learning from product warranty field data analysis. *2023 Annual Reliability and Maintainability Symposium (RAMS)*, 1–6. IEEE. https://doi.org/10.1109/RAMS51473.2023.10088234

Armbrust, M., Ghodsi, A., Xin, R., & Zaharia, M. (2021). Lakehouse: A New Generation of Open Platforms that Unify Data Warehousing and Advanced Analytics. *11th Annual Conference on Innovative Data Systems Research (CIDR '21)*. https://www.cidrdb.org/cidr2021/papers/cidr2021_paper17.pdf

Bach, B., Freeman, E., Abdul-Rahman, A., Turkay, C., Khan, S., Fan, Y., & Chen, M. (2023). Dashboard design patterns. *IEEE Transactions on Visualization and Computer Graphics, 29*(1), 342–352. https://doi.org/10.1109/TVCG.2022.3209448

Ballard, C., Farrell, D. M., Gupta, A., Mazuela, C., & Vohnik, S. (2006). Dimensional modeling: In a business intelligence environment. *IBM Redbooks*. https://www.redbooks.ibm.com/redbooks/pdfs/sg247138.pdf

Biswas, N., Sarkar, A., & Mondal, K. C. (2019). Efficient incremental loading in ETL processing for real-time data integration. *Innovations in Systems and Software Engineering, 16*(1), 53–61. https://doi.org/10.1007/s11334-019-00344-4

Blischke, W. (1993). Warranty Cost Analysis. *CRC Press.*

Blischke, W., Murthy, D. N. P. (1995). Product Warranty Handbook. *CRC Press.*

Borrohou, S., Fissoune, R., & Badir, H. (2025). The role of data transformation in modern analytics: A comprehensive survey. *Journal of Computer Languages*, *84*, 101329. https://doi.org/10.1016/j.cola.2025.101329

Chandra, P., & Gupta, M. K. (2018). Comprehensive survey on data warehousing research. *International Journal of Information Technology, 10*, 217–224. https://doi.org/10.1007/s41870-017-0067-y

Chaudhuri, S., & Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM SIGMOD Record*, *26*(1), 65–74. https://doi.org/10.1145/248603.248616

Databricks. (2026a). *The Databricks Data Intelligence Platform.* https://www.databricks.com/product/data-intelligence-platform

Databricks. (2026b). *Unity Catalog metric views.* https://docs.databricks.com/aws/en/metric-views/

Databricks. (2025). *What is the medallion lakehouse architecture?*. https://docs.databricks.com/gcp/en/lakehouse/medallion

Dmitriyev, V., Mahmoud, T., & Marín-Ortega, P. M. (2015). SOA enabled ELTA: Approach in designing business intelligence solutions in Era of Big Data. *International Journal of Information Systems and Project Management*, *3*(3), 49–63. https://doi.org/10.12821/ijispm030303

Doan, A., Halevy, A. & Ives, Z. (2012). *Principles of Data Integration*. Elsevier Science & Technology.

El-Sappagh, S. H. A., Hendawi, A. M. A., & El Bastawissy, A. H. (2011). A proposed model for data warehouse ETL processes. *Journal of King Saud University - Computer and Information Sciences*, *23*(2), 91–104. https://doi.org/10.1016/j.jksuci.2011.05.005

Fang, H. (2015). Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem. *2015 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, 820–824. https://doi.org/10.1109/CYBER.2015.7288049

Gaol, F. L., Siswanto, R. A., & Matsuo, T. (2023). Architectural modeling of data warehouse and analytic business intelligence for Bedstead manufacturers. *Open Engineering*, *13*(1). https://doi.org/10.1515/eng-2022-0508

Gaur, H., Sivakumar, D., Tao, T., Valani, Z., Fenton, J., & Sepp, T. (2025). *What's new with data sharing & collaboration.* Databricks. https://www.databricks.com/blog/whats-new-data-sharing-collaboration

Golfarelli, M., Rizzi, S., & Cella, I. (2004). Beyond data warehousing: What's next in business intelligence? *Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP, DOLAP '04*, 1–6. https://doi.org/10.1145/1031763.1031765

Hashemi-Pour, C., Scardina, J., & Horwitz, L. (2024). *What is Microsoft Power BI? Uses, features and guide*. Search Content Management. https://www.techtarget.com/searchcontentmanagement/definition/Microsoft-Power-BI

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design Science in Information Systems Research. *Management Information Systems Quarterly*, *28*, 75.

Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian Journal of Information Systems, 19*(2), Article 4. https://aisel.aisnet.org/sjis/vol19/iss2/4

Ivanković, D., Barbazza, E., Bos, V., Brito Fernandes, Ó., Jamieson Gilmore, K., Jansen, T., Kara, P., Larrain, N., Lu, S., Meza-Torres, B., Mulyanto, J., Poldrugovac, M., Rotar, A., Wang, S., Willmington, C., Yang, Y., Yelgezekova, Z., Allin, S., Klazinga, N., & Kringos, D. (2021). Features constituting actionable COVID-19 dashboards: Descriptive assessment and expert appraisal of 158 public web-based COVID-19 dashboards. *Journal of Medical Internet Research, 23*(2), e25682. https://doi.org/10.2196/25682

Jack, N., & Van Der Duyn Schouten, F. (2000). Optimal repair–replace strategies for a warranted product. *International Journal of Production Economics*, *67*(1), 95–100. https://doi.org/10.1016/S0925-5273(00)00012-8

Jiménez-Partearroyo, M., & Medina-López, A. (2024). Leveraging Business Intelligence Systems for Enhanced Corporate Competitiveness: Strategy and Evolution. *Systems*, *12*(3), 94. https://doi.org/10.3390/systems12030094

Jo, J., & Lee, K.-W. (2019). MapReduce-Based D_ELT Framework to Address the Challenges of Geospatial Big Data. *ISPRS International Journal of Geo-Information*, *8*(11), 475. https://doi.org/10.3390/ijgi8110475

Khan, B., Khan, W., Jan, S., & Chughtai, M. I. (2024). An Overview of ETL Techniques, Tools, Processes and Evaluations in Data Warehousing. *Journal on Big Data*, *6*(1), 1–20. https://doi.org/10.32604/jbd.2023.046223

Khan, M. A., Saqib, S., Alyas, T., Ur Rehman, A., Saeed, Y., Zeb, A., Zareei, M., & Mohamed, E. M. (2020). Effective Demand Forecasting Model Using Business Intelligence Empowered With Machine Learning. *IEEE Access*, *8*, 116013–116023. https://doi.org/10.1109/ACCESS.2020.3003790

King, T., & Schwarzenbach, J. (2020). *Managing Data Quality: A Practical Guide.* BCS Learning & Development Limited.

Kottek, M., Grieser, J., Beck, C., Rudolf, B., & Rubel, F. (2006). World map of the Köppen–Geiger climate classification updated. *Meteorol. Z., 15*, 259–263. https://doi.org/10.1127/0941-2948/2006/0130

Lawton, G. (2006). Making Business Intelligence More Useful. *Computer*, *39*(9), 14–16. https://doi.org/10.1109/MC.2006.318

Mahanti, R. (2019). *Data Quality: Dimensions, Measurement, Strategy, Management, and Governance.* Quality Press.

Mahlamäki, K., Niemi, A., Jokinen, J., & Borgman, J. (2016). Importance of maintenance data quality in extended warranty simulation. *International Journal of COMADEM*, *19*, 3–10.

Marshall, S., Arnold, R., Chukova, S., & Hayakawa, Y. (2018). Warranty cost analysis: Increasing warranty repair times. *Applied Stochastic Models in Business and Industry*, *34*(4), 544–561. https://doi.org/10.1002/asmb.2323

Microsoft. (2026). *What is Power BI? Overview of components and benefits - Power BI.* https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview

Miller, R., Whelan, H., Chrubasik, M., Whittaker, D., Duncan, P., & Gregório, J. (2024). A Framework for Current and New Data Quality Dimensions: An Overview. *Data*, *9*(12), 151. https://doi.org/10.3390/data9120151

Muntean, M., Dănăiaţă, D., Hurbean, L., & Jude, C. (2021). A Business Intelligence & Analytics Framework for Clean and Affordable Energy Data Analysis. *Sustainability*, *13*(2), 638. https://doi.org/10.3390/su13020638

Murthy, D. N. P. (2006). Product warranty and reliability. *Annals of Operations Research*, *143*(1), 133–146. https://doi.org/10.1007/s10479-006-7377-y

Negash, S., & Gray, P. (2008). Business Intelligence. *Springer,* 175–193. https://doi.org/10.1007/978-3-540-48716-6_9

Ojala, S. (2024). *Predictive Modelling and AI Integration for Enhanced Analysis of Warranty and Notification Data*. [Master's thesis, University of Vaasa]. Osuva. https://osuva.uwasa.fi/server/api/core/bitstreams/a8926438-cb6b-4483-8e71-25ed05f25024/content

Patel, H. (2025). *Field quality data visualization and fault prediction: A case study on data visualization at Scania.* [Master's thesis, Uppsala University]. Diva. https://www.diva-portal.org/smash/record.jsf?pid=diva2:2018320

Peffers, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *J. Manage. Inf. Syst.*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Ponniah, P. (2010). *Data Warehousing Fundamentals for IT Professionals, Second Edition.* Wiley.

Rayarao, S. R., & Donikena, N. (2025). Databricks Data Intelligence Platform: A comprehensive analysis of machine learning capabilities and data management features. *Authorea.* https://doi.org/10.22541/au.175709088.80309103/v1

Saket, B., Endert, A., & Demiralp, Ç. (2019). Task-Based Effectiveness of Basic Visualizations. *IEEE Transactions on Visualization and Computer Graphics*, *25*(7), 2505–2512. https://doi.org/10.1109/TVCG.2018.2829750

Schneider, J., Gröger, C., Lutsch, A., Schwarz, H., & Mitschang, B. (2024). The Lakehouse: State of the Art on Concepts and Technologies. *SN Computer Science*, *5*(5), 449. https://doi.org/10.1007/s42979-024-02737-0

Sedrakyan, G., Mannens, E., & Verbert, K. (2019). Guiding the choice of learning dashboard visualizations: Linking dashboard design and data visualization concepts. *Journal of Computer Languages*, *50*, 19–38. https://doi.org/10.1016/j.jvlc.2018.11.002

Sharda, R., Delen, D., & Turban, E. (2018). *Business intelligence, analytics, and data science: A managerial perspective (Fourth edition).* Pearson.

Sherman, R. (2014). *Business Intelligence Guidebook: From Data Integration to Analytics*. Elsevier Science & Technology.

Silva, D., Paulo, P., & Amaro, A. (2020). Logistic Performance & Dashboards: A flexible Power BI solution. *CAPSI 2020 Proceedings*. https://aisel.aisnet.org/capsi2020/1

Simitsis, A., Skiadopoulos, S., & Vassiliadis, P. (2023). The History, Present, and Future of ETL Technology (invited). *CEUR Workshop Proceedings.* https://www.semanticscholar.org/paper/The-History%2C-Present%2C-and-Future-of-ETL-Technology-Simitsis-Skiadopoulos/017ed88c7a7f4d72319bcf5821b591b694fd23f7

Sinha, C. (2021). *Mastering Power BI: Build Business Intelligence Applications Powered with DAX Calculations, Insightful Visualizations, Advanced BI Techniques, and Loads of Data Sources.* BPB Publications.

Souibgui, M., Atigui, F., Zammali, S., Cherfi. S. & Yahia, S. (2019). Data quality in ETL process: A preliminary study. *Procedia Computer Science.* https://doi-org.proxy.uwasa.fi/10.1016/j.procs.2019.09.223

Sreedhar, T. S., Islam, S., Atmosa, M., Yazdandoust, E., Elnaim, M. S., Mishra, S., Vemparala, V. N., & Bajpai, R. (2024). Applications of big data in renewable energy systems based on cloud computing. *International Journal on Information Technologies & Security, 16*(3), 121–131. https://ijits-bg.com/sites/default/files/archive/2024%28vol.16%29/No3/contents/2024-N3-12.pdf

Srinivasan, A., Drucker, S. M., Endert, A., & Stasko, J. (2019). Augmenting Visualizations with Interactive Data Facts to Facilitate Interpretation and Communication. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 672–681. https://doi.org/10.1109/TVCG.2018.2865145

Storey, V. C., & Song, I.-Y. (2017). Big data technologies and Management: What conceptual modeling can do. *Data & Knowledge Engineering, 108,* 50–67. https://doi.org/10.1016/j.datak.2017.01.001

Sundar, D., Jayaram, Y., & Bhat, J. (2022). A Comprehensive Cloud Data Lakehouse Adoption Strategy for Scalable Enterprise Analytics. *International Journal of Emerging*

Research in Engineering and Technology, *3*(4), 92–103. https://doi.org/10.63282/3050-922X.IJERET-V3I4P111

Thuan, N., Drechsler, A., & Antunes, P. (2019). Construction of Design Science Research Questions. *Communications of the Association for Information Systems*, *44*(1). https://doi.org/10.17705/1CAIS.04420

Vassiliadis, P. (2009). A Survey of Extract-Transform-Load Technology. *International Journal of Data Warehousing and Mining.* https://doi.org/10.4018/jdwm.2009070101

Vintr, Z., & Vintr, M. (2007). Estimate of Warranty Costs Based on Research of the Customer's Behavior. *2007 Annual Reliability and Maintainability Symposium*, 323–328. https://doi.org/10.1109/RAMS.2007.328135

vom Brocke, J., Hevner, A., & Maedche, A. (2020). *Design Science Research. Cases*. Springer International Publishing AG.

Walha, A., Ghozzi, F., & Gargouri, F. (2024). Data integration from traditional to big data: Main features and comparisons of ETL approaches. *The Journal of Supercomputing*, *80*(19), 26687–26725. https://doi.org/10.1007/s11227-024-06413-1

Wallace, R., Blischke, M., Rezaul, K. & Murthy, D. N. P. (2011). *Warranty Data Collection and Analysis*. Springer Science & Business Media.

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an information system design theory for vigilant EIS. *MIS Quarterly, 16*(1), 36–58. http://www.jstor.org/stable/23010780.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *JMIS. Journal of Management Information Systems, 12*(4), 5. https://www.proquest.com/scholarly-journals/beyond-accuracy-what-data-quality-means-consumers/docview/218911948/se-2

Wang, X., He, K., He, Z., Li, L., & Xie, M. (2019). Cost analysis of a piece-wise renewing free replacement warranty policy. *Computers & Industrial Engineering*, *135*, 1047–1062. https://doi.org/10.1016/j.cie.2019.07.015

Wieder, B., & Ossimitz, M.-L. (2015). The Impact of Business Intelligence on the Quality of Decision Making – A Mediation Model. *Procedia Computer Science, Conference*

on Enterprise Information Systems/International Conference on Project Management/Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN / HCist 2015 October 7-9, *64*, 1163–1171. https://doi.org/10.1016/j.procs.2015.08.599

Winter, R. (2008). Design science research in Europe. *European Journal of Information Systems, 17*, 470–475. https://doi.org/10.1057/ejis.2008.44

Wu, S. (2012). Warranty Data Analysis: A Review. *Quality and Reliability Engineering International*, *28*(8), 795–805. https://doi.org/10.1002/qre.1282

Wu, S. (2013). A review on coarse warranty data and analysis. *Reliability Engineering & System Safety*, *114*, 1–11. https://doi.org/10.1016/j.ress.2012.12.021

Yu, X., Bryant, C., Wheeler, N. R., Rubel, F., Ascencio Vasquez, J., & French, R. H. (2023). *kgcPy: Koeppen-Geiger climatic zones documentation*. kgcPy. https://kgcpy.readthedocs.io/en/latest/index.html

Zhang, N., Zhang, J., Jiang, S., & Ge, W. (2024). The Effects of Layout Order on Interface Complexity: An Eye-Tracking Study for Dashboard Design. *Sensors, 24*(18), 5966. https://doi.org/10.3390/s24185966

Appendices

Appendix 1. Use of artificial intelligence

Artificial intelligence (AI) tools were used to support the writing process by correcting grammatical errors, improving the fluency of the text, and assisting in structuring paragraphs based on the content provided by the author. All corrections and suggestions generated by AI were evaluated, modified if needed, and approved by the author. AI was used only for supportive tasks, and all research activities, and interpretations were conducted independently by the author. The AI application utilized in this study was Microsoft Copilot.