Thesis for The Degree of Doctor of Philosophy

Synergizing Data Management, DataOps, and Data Pipelines for AI-Enhanced Embedded Systems

Aiswarya Raj Munappy

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2024

Synergizing Data Management, DataOps, and Data Pipelines for AI-Enhanced Embedded Systems

Aiswarya Raj Munappy

ISBN 978-91-8103-052-5
Doktorsavhandlingar vid Chalmers tekniska högskola, Ny serie nr 5510.
ISSN 0346-718X
Technical Report No. 195D

Department of Computer Science and Engineering
Division of Software Engineering
AI Engineering Research Group
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2024.

"To my parents, who never imagined their bedtime stories would inspire a thesis; to my brother, for keeping me grounded with much-needed humor; to my husband, who mastered the art of nodding and pretending to understand my research babble; and to my daughter, the pint-sized philosopher for reminding me that anything is possible—even this thesis!"

Synergizing Data Management, DataOps, and Data Pipelines for AI-Enhanced Embedded Systems

Aiswarya Raj Munappy

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

Context: Data management is a critical aspect of any artificial intelligence (AI) initiative, playing a pivotal role in the development, training, and deployment of AI models. A well-structured approach to data management ensures that AI models are trained on reliable data, comply with ethical standards, and contribute positively to decision-making processes in embedded systems.
Objectives: This thesis is structured around three primary objectives. The first objective is to comprehensively understand and address the data management challenges associated with embedded systems. Building upon this understanding, the second objective is to explore the data management practices that can help alleviate the challenges of data management. Finally, the third objective aims to develop and validate the implementation approaches for enhanced data management.
Method: To achieve the objectives, we conducted research in close collaboration with industry and used a combination of different empirical research methods like interpretive case studies, literature reviews, and action research.
Results: This thesis presents six main results. First, it identifies and categorizes data management challenges, solutions, and limitations. Second, it presents a stairway model delineating the stages of the evolution towards DataOps. Third, it proposes a model for evaluating the maturity of data pipelines and identifies determinants to assess the impact of machine learning (ML) on data pipelines. Fourth, it identifies the differences between unidirectional and bidirectional data pipelines and the significance, benefits, and challenges of bidirectional data pipelines. The thesis also provides a roadmap for the smooth migration from unidirectional to bidirectional data pipelines. Fifth, it presents and validates the conceptual model of an end-to-end data pipeline for ML/DL models. Finally, it presents and validates fault-tolerant data pipelines and an AI-powered 4-stage model for automated fault recovery in data pipelines.
Conclusion: In conclusion, this thesis demonstrates a well-structured approach to data management in AI-enhanced embedded systems, supported by innovative practices and robust implementation approaches, that is essential for ensuring the reliability, and effectiveness of data in decision-making processes.

Keywords

Data Management Challenges, DataOps Evolution, Data Pipelines, Robustness, Bidirectional, Fault-Tolerance, Automated Fault Recovery

List of Publications

Included publications

This thesis is based on the following publications:

[**A**] **Munappy, Aiswarya Raj**, Bosch, Jan and Olsson, Helena Holmström and Arpteg, Anders and Brinne, Björn, *Data management for production quality deep learning models: Challenges and solutions*
Journal of Systems and Software (2022).

[**B**] **Munappy, Aiswarya Raj**, Mattos, D. I., Bosch, J., Olsson, H. H., & Dakkak, A., *From ad-hoc data analytics to dataops*
In Proceedings of the International Conference on Software and System Processes (165-174) (2020, June).

[**C**] **Munappy, Aiswarya Raj**, Bosch, J., Olsson, H. H., & Jansson, A., *On the Impact of ML use cases on Industrial Data Pipelines*
In 2021 28th Asia-Pacific Software Engineering Conference (APSEC) (pp. 463-472) (2021, December) IEEE.

[**D**] **Munappy, Aiswarya Raj**, Bosch, J., & Olsson, H. H., *Data pipeline management in practice: Challenges and opportunities*
In Product-Focused Software Process Improvement: 21st International Conference, PROFES 2020, Turin, Italy, November 25–27, 2020, Proceedings 21 (pp. 168-184). Springer International Publishing.

[**E**] **Munappy, Aiswarya Raj**, Jan Bosch, Helena Holmström Olsson, *Maturity Assessment Model for Industrial Data Pipelines*
In 2023 30th Asia-Pacific Software Engineering Conference (APSEC) (pp. 463-472) (2023, December) IEEE.

[**F**] **Munappy, Aiswarya Raj**, Jan Bosch, Helena Holmström Olsson, & Dakkak, A., *Bidirectional Data Pipelines: An Industrial Case Study*
Submitted, under review.

[**G**] **Munappy, Aiswarya Raj**, Jan Bosch, Helena Holmström Olsson, and Tian J. Wang, *Modelling data pipelines*
In 2020 46th Euromicro conference on software engineering and advanced applications (SEAA), pp. 13-20. IEEE, 2020.

[**H**] **Munappy, Aiswarya Raj**, Jan Bosch, Helena Holmström Olsson, *On the Trade-off Between Robustness and Complexity in Data Pipelines*
In Quality of Information and Communications Technology: 14th International Conference, QUATIC 2021, Algarve, Portugal, September 8–11, 2021, Proceedings 14 (pp. 401-415). Springer International Publishing.

[**I**] **Munappy, Aiswarya Raj**, Jan Bosch, Helena Holmström Olsson, and Tian J. Wang, *Towards Automated Detection of Data Pipeline Faults*
In 2020 27th Asia-Pacific Software Engineering Conference (APSEC), pp. 346-355. IEEE, 2020.

[**J**] **Munappy, Aiswarya Raj**, Jan Bosch, Helena Holmström Olsson, *AI Powered Fault tolerance in Data Pipelines*
Submitted, under review.

Other publications

The following publications were published during my PhD studies, or are currently in submission/under revision. However, they are not appended to this thesis, due to contents not related to the thesis.

[**K**] **Munappy, Aiswarya Raj**, Jan Bosch, Helena Holmström Olsson, Anders Arpteg, and Björn Brinne, *Data management challenges for deep learning*
In 2019 45th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), pp. 140-147. IEEE, 2019..

[**L**] Lwakatare, Lucy Ellen, **Munappy, Aiswarya Raj**, Jan Bosch, Helena Holmström Olsson, and Ivica Crnkovic, *A taxonomy of software engineering challenges for machine learning systems: An empirical investigation. In Agile Processes in Software Engineering and Extreme Programming: 20th International Conference, XP 2019, Montréal, QC, Canada, May 21–25, 2019, Proceedings 20, pp. 227-243. Springer International Publishing, 2019..*

[**M**] Lwakatare, Lucy Ellen, **Munappy, Aiswarya Raj**, Crnkovic, I., Jan Bosch, Helena Holmström Olsson, *Large-scale machine learning systems in real-world industrial settings: A review of challenges and solutions. (2020). Information and software technology, 127, 106368..*

[**N**] Dakkak, A., **Munappy, Aiswarya Raj**, Jan Bosch, Helena Holmström Olsson, *Customer Support In The Era of Continuous Deployment: A Software-Intensive Embedded Systems Case Study*
In 2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC) (pp. 914-923). IEEE.

Acknowledgment

First and foremost, I extend my deepest gratitude to my main supervisor, Jan Bosch, for his unwavering support during my Ph.D. journey. His mentorship, profound knowledge, and motivational guidance have been pivotal in my research progress. His valuable feedback and advice have significantly refined my analytical skills and elevated my work. I am truly fortunate to have had him as my advisor and mentor. Additionally, I am thankful to my co-advisor, Helena Holmström Olsson, for her significant contributions. Specifically, her expertise was crucial in developing the research questions and methodology, and her guidance on interview techniques, data analysis, and her consistent availability and positivity laid a solid foundation for my qualitative research. I am also grateful to all the staff at Chalmers University, especially to my former examiner, the late Ivica Crnkovic, for his priceless guidance, and to my current examiner, Miroslaw Staron, for his responsible commitment and dedication. Further, I am grateful to Eric Knauss and Jennifer Horkoff for managing my teaching responsibilities effectively. My gratitude also goes to Lucy Ellen Lwakatare, David Issa Mattos, Teodor Fredriksson, and Meenu Mary John for the valuable learning experiences during our interactions. Additionally, I thank Anas Dakkak, Lotte Ansgaard Thomson, Emir Karamehmetoglu, Damir Basic Knezevic, Anders Jansson, and Tian J. Wang for their contributions to my research, making it a stimulating experience beyond the office environment. Last but not least, I am immensely thankful to the Software Center companies and the practitioners for their collaboration and support throughout my research. Their cooperation and encouragement were crucial for the success of this project.

Contents

Chapter 1

Introduction

Artificial Intelligence (AI) has transitioned from a niche field to a ubiquitous and indispensable technology, playing a transformative role across various industry domains. AI has evolved as a horizontal technology and serves as a foundational layer that enhances functionalities and augments capabilities across diverse fields. Unprecedented availability of data and advancements in high-performance parallel hardware, such as GPUs and FPGAs are the two driving forces behind this evolution which in turn paved the way for the rapid adoption of ML/DL solutions, with virtually every company now engaging in AI initiatives. However, research [1] [2] [3] [4] reveals that organizations still face significant challenges in transitioning from prototype to deployment of production-quality AI models. One of the major reasons is the lack of attention given to the vast and complex required surrounding infrastructure of AI systems as stated by Sculley, David, et al. [5]. The key challenges companies face during AI adoption relate to data quality, design methods, model performance, deployment, and compliance [6].

Data being a critical factor, efficient data management is essential for the successful development and deployment of AI-enhanced embedded systems. Effective data management ensures that the right data is collected, stored, processed, and utilized efficiently in the organization [7]. Without proper data management, organizations encounter challenges such as data inconsistency, inaccuracies, and inefficiencies, leading to degraded performance and unreliable outcomes [4]. Additionally, robust data management practices facilitate model training, optimization, and adaptation, enabling embedded systems to continuously learn and improve their performance over time [8].

AI-enhanced embedded systems refer to hardware and software components that integrate artificial intelligence capabilities into embedded devices, enabling them to perform advanced tasks, optimize performance, and provide enhanced functionality, ranging from predictive maintenance and autonomous navigation to personalized user experiences and smart automation. Data management is particularly important in this domain due to the following reasons. Many embedded systems operate in real-time or near-real-time environments, where timely processing of data is essential. Efficient data management enables quick retrieval, analysis, and response to incoming data streams, ensuring that AI models can make decisions or take actions within stringent time constraints [9]. Further, proper data management practices, such as data validation, cleaning, and normalization, help maintain data quality and integrity, ensuring that AI-enhanced embedded systems produce accurate and reliable results [10]. Furthermore, effective data management facilitates continuous learning and adaptation by providing mechanisms for updating models, incorporating new data, and refining algorithms based on feedback from the environment [11].

The challenges of data management in production machine learning are multifaceted. Experts at Google highlight the importance of robust processes for analyzing, validating, and transforming data fed into ML systems emphasizing the constraints imposed at different stages of the model's lifecycle [12]. They published another study that discusses challenges related to data understanding, validation, cleaning, and preparation in large-scale machine-learning systems [13]. Bhowmik et. al [14] also highlight the need for AI research to focus on the data-centric approach alongside the model-centric approach, exploring and developing methodologies for improving data quality, consistency, labeling, and performance auditing and investigating the impact of data-centric approaches on improving the accuracy and performance of ML models.

While research in the broader context of data management and artificial intelligence is abundant, there exists a paucity of studies focusing specifically on the best practices tailored to mitigate the data management challenges. Similarly, strategies for implementing efficient data management practices within the constraints of embedded systems architecture are limited. Furthermore, implementation strategies for overcoming evolving data quality problems, such as incorporating machine learning algorithms to dynamically adjust data processing techniques are relatively unexplored. Thus, the exploration of data management challenges and corresponding strategies tailored to the context of AI-enhanced embedded systems represents a fertile area for research and innovation.

Mitigation strategies in the context of data management challenges for AI-enhanced embedded systems involve implementing solutions or best practices to prevent or reduce the impact of these challenges. One such approach is DataOps, which is a methodology that combines data engineering, data integration, and data quality processes to improve the flow of data between data sources and data consumers. By implementing DataOps practices, organizations can streamline data management processes, ensure data quality, and enhance collaboration between data engineers, data scientists, and other stakeholders involved in the data pipeline.

Implementation strategies, on the other hand, refer to the ability of data pipelines to dynamically adjust to changing conditions, requirements, or constraints. In the context of AI-enhanced embedded systems, implementation strategies involve designing data pipelines that can efficiently handle evolving data sources, changing model requirements, and dynamic system conditions. This may include incorporating techniques such as automated fault-tolerance to ensure that the data pipeline can adapt to new challenges and requirements as they arise.

This thesis analyzes best practices and implementation strategies that are essential for alleviating data management challenges in AI-enhanced embedded systems. Best practices such as DataOps and data pipelines can help organizations proactively address data quality issues, streamline data processes, and improve collaboration among data stakeholders. On the other hand, implementation approaches discussed in the thesis ensure that data pipelines remain flexible and responsive to changing requirements, enabling organizations to effectively manage dynamic data sources, evolving model needs, and shifting system conditions. By adopting practices for mitigating data management challenges and implementing approaches for improved data management, organizations can enhance the efficiency, reliability, and effectiveness of their data management processes in the context of AI-enhanced embedded systems.

The thesis was conducted in the context of the Software Center and emphasized the empirical nature of the research by conducting collaborative studies with industry partners to address specific challenges faced by them. Through close collaboration with four different companies globally, the research focused on real-world problems relevant to the industry partners. The included and related publications were a result of discussions with these industrial partners, highlighting the practical and empirical approach taken in the research process.

The remainder of this thesis is structured as follows: Chapter 2 provides a background review of the main concepts utilized throughout this thesis. In Chapter 3, the objectives of this thesis, the research questions, and an overview of the different research strategies and data analysis methods used in the included publications are presented. Chapter 4 discusses each included publication, providing a summary of its main contributions and how they relate to publications produced in this thesis. Chapters 5 to 14 contain the included publications. Chapter 15 discusses the proposed objectives and research questions in light of the included publications. Finally, Chapter 16 concludes this thesis and discusses potential research directions for data management in AI-enhanced embedded systems.

Chapter 2

Background

Data management serves as a cornerstone that shapes the development, training, and deployment of AI models. Particularly in the context of embedded systems, where AI plays a pivotal role in driving technological innovations, the effective management of data is crucial to ensure the reliability, ethical compliance, and positive impact of AI models on decision-making processes. This thesis studies data management challenges, data management practices, and implementation approaches for improved data management specifically for AI-enhanced embedded systems. It caters to a diverse audience, including AI researchers, developers, industry professionals, and stakeholders seeking to deepen their understanding of the intricate interplay between data management and AI in embedded systems. This section provides a review of contemporary literature relevant to better understand the remainder of the thesis.

2.1 Rise of Data Complexity

The exponential growth of data volume, variety, and velocity in modern organizations, often referred to as the three Vs of big data (Volume, Variety, Velocity), presents both opportunities and challenges for businesses across industries [15]. There are several challenges posed by managing and extracting value from large and diverse datasets.

2.1.1 Data Volume:

The sheer volume of data generated and collected by organizations is increasing exponentially, with data being produced at an unprecedented rate. Managing and storing massive amounts of data can quickly become overwhelming, leading to issues related to scalability, storage costs, and data redundancy [16]. Traditional data management systems struggle to efficiently process and analyze large volumes of data within acceptable time frames, hindering real-time decision-making and insights generation. Extracting valuable insights from vast datasets requires advanced analytics tools and techniques that can handle big data processing efficiently and effectively [16].

2.1.2 Data Variety:

Data in modern organizations comes in various formats, including structured data from databases, unstructured data from social media, text, images, videos, and semi-structured data like log files [17]. Managing diverse data types can be complex and challenging. Integrating and harmonizing disparate data sources to derive meaningful insights and maintain data quality poses challenges related to data governance, data integration, and data quality assurance [18] [17]. Traditional data processing methods may not be equipped to handle the diversity of data formats and sources, leading to siloed data and missed opportunities for holistic analysis.

2.1.3 Data Velocity:

Data is being generated and updated at an unprecedented speed, requiring organizations to process and analyze data in near real-time to extract timely insights and respond to dynamic business needs. Managing high-velocity data streams, such as IoT sensor data, financial transactions, or social media feeds, poses challenges in terms of data ingestion, processing speed, and ensuring data freshness [19] [20]. Traditional batch processing approaches may not be suitable for handling high-velocity data streams, necessitating the adoption of real-time data processing solutions like stream processing frameworks [21].

In conclusion, the exponential growth of data volume, variety, and velocity in modern organizations presents significant challenges in terms of data management, processing, and deriving value from large and diverse datasets. To address these challenges effectively, organizations need to invest in scalable infrastructure, advanced analytics capabilities, data integration technologies, and agile data processing frameworks that can handle the complexities of big data and enable actionable insights for informed decision-making.

2.2 Synergy between Data and Artificial Intelligence

AI's ability to work well with data analytics is the main reason for data being an integral part of AI. AI algorithms like machine learning and deep learning are capable of mining every small detail from the input data and those inputs are used to generate new rules to fulfill its function [22]. Data and AI are merging into a synergistic relationship, where AI is useless without data and data is insurmountable without AI. Big Data will continue to grow larger as AI becomes a viable option for automating more activities, and AI will become a bigger field as more data is available for learning and analysis. Moreover, business decisions are based on big data that previously were based on guesswork or painstakingly constructed models of reality [23]. The sheer volume and variety of data consumed by modern analytical pipelines have greatly strengthened the connections between data integration and machine learning [24]. Data management systems are increasingly using AI models like machine learning to automate parts of data life cycle tasks. Examples include data cataloging and inferring the schema of raw data [25]. Data analytics drives nearly every aspect of our modern society, including mobile services, retail manufacturing, financial services, life sciences, and physical sciences [23]. In most industries, established competitors and new entrants alike will leverage data-driven strategies to innovate, compete, and capture value from deep and up to real-time information [26]. However, in the current scenario organizations struggle with collecting, integrating, and managing the data. AI will not solve these data issues, rather it will only make them more noticeable.

2.3 AI-enhanced Embedded Systems

Currently, considerable transitions are ongoing in the embedded systems industry, i.e. markets becoming more fast-changing and unpredictable, customer requirements becoming increasingly complex, rapidly advancing technologies, and the constant need to shorten the time-to-market of new products [27]. Moreover, while the ability to manufacture high-quality mechanical subsystems remains perilous, it is no longer the key identifier and what makes a company competitive. During the last decade, along with electronics and software, AI has been introduced into many products, and embedded systems companies are becoming increasingly AI-driven [28]. AI/ML is becoming a horizontal technology: its application is expanding to more domains. Embedded Systems are also increasingly integrating AI into applications for performance improvement [29] [30]. Applications that involve both “traditional” software and Artificial Intelligence components are referred to as AI-enhanced Embedded Systems throughout this thesis. For instance, an embedded system that uses sensors to monitor things like temperature and vibration. Such a system should be able to detect anomalies in the early stages of things starting to go wrong, make predictions about future events, and alert its human supervisors as to what’s going on. Here, AI is not the key component that controls the whole system, but it is used to enhance the performance of the entire system. Since AI-enhanced embedded systems rely heavily on software, it is expected that Software Engineering methods and tools can help. However, the development differs from the development of “traditional” software systems in a few substantial aspects. Hence, traditional SE methods and tools are not sufficient by themselves and need to be adapted and extended. AI-enhanced applications and AI-intensive applications are very common in the online domain. However, in the Embedded System domain mechanical subsystems, electronics, and software are integral parts of embedded systems. Consequently, the developers won’t be experts in AI application development, which in turn makes integration of AI components difficult. Moreover, the data will be generated by both software and AI components. Thus, volume, velocity, and variety of data increase and should be managed accordingly to reap maximum benefits from the data.

2.4 Data management for AI-enhanced Embedded Systems

Data management is an administrative process that includes acquiring, validating, storing, protecting, and processing required data to ensure the accessibility, reliability, and timeliness of the data [31]. Inappropriate treatment of data leads to data becoming corrupt, unusable, or completely useless. Companies trying to become data-driven are increasingly collecting and storing data from all possible sources. However, such companies need to understand that simply collecting data is not enough instead there is the need to understand from the start that data management and data analytics will be successful only after putting some insights into how to gain value from the collected raw data [22]. Efficient systems for processing, storing, and validating data, as well as effective analysis strategies, are required beyond data collection. Each step of data collection and management must lead towards acquiring the right data and analyzing it to get the actionable intelligence that is required to make data-driven decisions [32]. Managing the data is the first step towards handling the large volume of data, both structured and unstructured, online and offline, that floods daily. Data management best practices enable organizations to harness the full power of the data and gain the insights needed to make the data useful [33].

When designing artificial intelligence solutions, practitioners spend a significant amount of time focusing on aspects such as the nature of the problem, selection of learning algorithms, etc. However, little attention is often provided to the data on which the AI solution operates. As it turns out, the characteristics of the data are one of the absolute key elements that determine the right models for an AI solution. One possible reason for this indifference is that significant research has been done on data management practices over years. However, data required for AI models need to undergo substantial pre-processing before feeding it to the models. Moreover, the volume, variety, velocity, and veracity of data are increasing daily which acts as a compelling reason for the development of data management practices specifically for AI models.

2.5 Demand for Agile Data Operations and Adoption of DataOps Principles

The exponential growth of digital data from various sources such as sensors and devices has led to challenges in analyzing and deriving useful insights from this vast amount of information. Traditional approaches to data management struggle to keep up with the high velocity of data and the demands of real-time analytics, resulting in poor data quality and compromised trust in the data [34]. Companies are increasingly focusing on storing and processing large volumes of data to not only process it but also to derive accurate and timely conclusions [35]. To address these challenges, a systematic approach similar to Software Development Life Cycle (SDLC) is needed for the development of data products, taking into account the unique characteristics of big data and the available infrastructures, tools, and development models.

In response to the need for more agile and efficient data management practices, the principles of Agile Development have been applied to analytics development. By adopting an agile philosophy, data teams can share results more frequently, gather stakeholder feedback, and use that feedback to validate and evolve the analytics development towards an agreeable end-state [36]. This iterative and collaborative approach helps in dealing with the experimental nature of analytics development, where detailed requirements cannot always be predefined with complete confidence. By fostering a continuous feedback loop and adapting to evolving requirements, agile practices in analytics development aim to enhance the quality and relevance of the results generated.

Moreover, the emergence of DataOps as a new and independent approach to data analytics has gained traction in the industry. DataOps encompasses a set of practices aimed at bringing speed and agility to end-to-end data pipeline processes, from data collection to delivery [37]. By introducing Agile Development principles into data analytics, DataOps promotes collaboration and innovation within data teams, enabling more efficient and effective work processes. DataOps focuses on automating, orchestrating, and monitoring the flow of data through operations, with an emphasis on the value pipeline and the innovation pipeline. The value pipeline processes data to create insights or value, while the innovation pipeline introduces new insights or value into the data flow, akin to the DevOps framework [38].

In conclusion, the demand for agile data operations and the adoption of DataOps principles stems from the need to address the challenges posed by the exponential growth of data and the requirements for real-time analytics. By integrating agile methodologies and DataOps practices into data management processes, organizations can enhance their ability to derive valuable insights from data, improve data quality, and adapt more effectively to changing business requirements.

2.6 Ensuring Reliability in Data Pipelines

The management of data is best captured using its data pipeline. A data pipeline is a set of tools and activities for moving data from one system with its method of data storage and processing to another system in which it can be stored and managed differently. Moreover, pipelines allow for automatically getting data from many disparate sources, then transforming and consolidating it in one high-performing data storage [39]. Data Pipelines are a chain of activities that are connected, where each activity represents an atomic data task. Developing data pipelines enables the automation of most of the tasks in the data lifecycle. A data pipeline can be a simple process of data extraction and loading, or, it can be designed to handle data in a more advanced manner, such as training datasets for machine learning. Data pipelines are highly beneficial as they can process data in multiple formats from distributed data sources with minimal human intervention, accelerate data life cycle activities, and enhance productivity in data-driven enterprises [40]. Data pipelines enable traceability, and fault-tolerance, and reduce human errors through maximizing automation, thereby producing high-quality data [41]. However, a powerful argument against constructing a data pipeline is the cost of building and maintaining it, in terms of time, money, morale, and lost opportunities. Building a data pipeline demands specialized skills, time, and extensive experience in data engineering. Data pipeline construction is a task for which most data scientists have limited aptitude, interest, or training. Approximately 80% of an average data scientist's time is spent constructing data pipelines [42]. An alternate option is to buy a ready-made data pipeline from external vendors. As the use cases, organization culture, the expertise of the employees, etc. varies from one company to another, it is always better to design a tailor-made data pipeline that can meet the requirements of the company. Automated data pipelines allow simple and flexible integrations, pipeline transparency, and automated workflows and processes to support even the most aggressive data management plans, thereby delivering flexibility, scale, and cost-effectiveness.

Ensuring data pipeline reliability is critical for mission-critical applications in modern organizations, as any disruption or failure in the data processing workflow can have significant repercussions on business operations and decision-making processes. Reliable data pipelines ensure the continuous flow of data from source to destination without interruptions or delays [43] [44]. This is crucial for maintaining up-to-date information for decision-making processes. Further, they guarantee data consistency and accuracy throughout the processing stages, preventing data discrepancies and ensuring the integrity of analytical results [44]. Data pipelines also enable the timely delivery of insights and reports to stakeholders, supporting agile decision-making and operational efficiency. Pipeline failures can disrupt critical business operations, leading to delays in reporting, decision-making, and customer service delivery. In mission-critical applications, pipeline failures can result in financial losses due to missed opportunities, inaccurate insights, or operational inefficiencies. Persistent pipeline failures can tarnish the organization's reputation, eroding trust among customers, partners, and stakeholders [45]. Therefore, we need reliable data pipelines to ensure that decision-makers have access to accurate and timely data, enabling them to make informed decisions based on trustworthy insights. To increase the reliability of data pipelines, there is a need to improve operational efficiency by reducing manual intervention, minimizing downtime, and optimizing data processing workflows. Reliable data pipelines can also enhance business agility by enabling quick recovery from failures and adapting to changing data processing requirements [46].

In summary, the criticality of ensuring data pipeline reliability in mission-critical applications cannot be overstated, as the impact of pipeline failures can be detrimental to business operations, decision-making processes, and overall organizational performance. By prioritizing reliability and fault tolerance in data pipeline design and implementation, organizations can mitigate risks, ensure data integrity, and maintain operational continuity.

Chapter 3

Research Approach

In this chapter, we present the three main research objectives that contribute to the aim of this thesis, the particular research questions for the objective, and the research strategies and methods employed in the included publications.

3.1 Objectives

Each objective presented in this chapter provides clarity on the scope and purpose of this research. Fig. 3.1 illustrates the overarching aim of the study, and the mapping between the research objectives and the included publications.

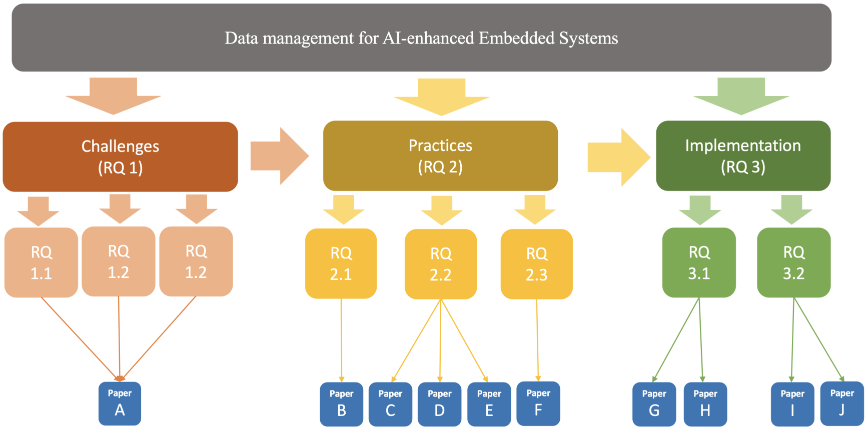


Figure 3.1: Research Overview

3.1.1 Objective 1: Identify the data management challenges

The first objective of this research was to identify the data management challenges throughout the data life cycle phases, enabling researchers to develop targeted solutions that address the specific needs and complexities of each stage of the data management process.

Identifying the data management challenges is foundational for our research, as it forms the initial step towards understanding the complexities of data management for AI-enhanced Embedded Systems. By gaining insights into these challenges, the study focuses on finding existing solutions, prevention strategies to address them, and the limitations of the solutions. Also, a comprehensive understanding of these challenges is vital for evaluating the current state of data management practices within AI-enhanced embedded systems. Moreover, the identification of data management challenges provides valuable insights for the data stakeholders so that resources can be allocated wisely to improve data management. Therefore, this objective is the foundation for the subsequent stages of the research, guiding the exploration and analysis of data management practices and implementation approaches aimed at improving data management in AI-enhanced embedded systems.

For this objective, the following research questions are studied in the included publication:

- **RQ1:** What are the data management challenges, limitations of existing solutions, and open research problems in the field?
  - **RQ 1.1.** What are the data management challenges experienced in industry?
  - **RQ1.2.** What are the inherent limitations of current solutions proposed to address the data management challenges?
  - **RQ1.3.** What persists as unresolved challenges in the field of data management?

This thesis investigates the challenges related to data management, solutions, and limitations in paper A discussed in Chapter 5.

3.1.2 Objective 2: Explore the data management practices that can alleviate data management challenges

The second objective entails analyzing the role and impact of DataOps and data pipeline management in revolutionizing data management practices, thereby paving the way for more agile, efficient, and resilient data ecosystems in AI-enhanced embedded systems. Data management challenges underscore the need for innovative approaches that can enhance the efficiency, reliability, and agility of data management practices. By delving into the details of these practices, the study aims to explain the practices, benefits, and implications of their application in the AI-enhanced Embedded Systems domain.

This thesis explores the role of DataOps in Chapter 6, analyzes the role of ML on the need for new data management practices in Chapter 7, discusses the challenges and opportunities in Chapter 8, and identifies the maturity stages of data pipelines in Chapter 9. For this objective, the following research questions and sub-questions are discussed in the included publications:

- **RQ2:** How can DataOps and Data pipelines alleviate the data management challenges in the context of AI-enhanced Embedded Systems?
  - **RQ2.1.** What specific contributions does the implementation of DataOps make to address and enhance data management practices, within the context of AI-enhanced embedded systems?
  - **RQ2.2.** How does the implementation of data pipelines contribute to addressing data management challenges, particularly within the context of AI-enhanced embedded systems?
  - **RQ2.3.** What role do bidirectional data pipelines play in enhancing data management practices, particularly within the domain of AI-enhanced embedded systems?

3.1.3 Objective 3: Investigate the implementation approaches for improved data management

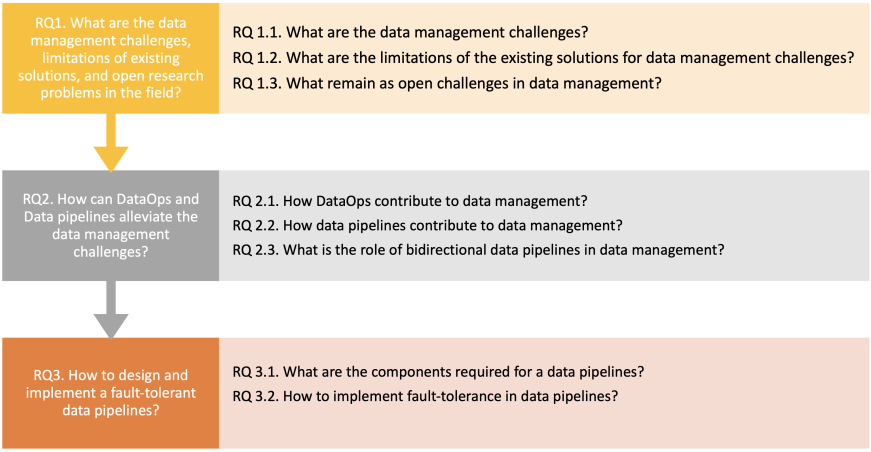
The third objective of our study is to explore the design and implementation of fault-tolerant data pipelines, focusing on identifying the essential components required for building robust, automated, and traceable end-to-end data pipelines in AI-enhanced Embedded Systems for improved data management. The study also seeks to investigate strategies for implementing fault tolerance in data pipelines and automating fault recovery mechanisms using artificial intelligence (AI).

The investigation of fault-tolerant data pipeline design and automation is significant for data management for several reasons. Firstly, by understanding the essential components and strategies for building fault-tolerant data pipelines, organizations can ensure the reliability and integrity of their data processing workflows, reducing the risk of data loss or inconsistencies due to failures. Moreover, the implementation of fault tolerance mechanisms and automation in data pipelines can lead to increased operational efficiency and reduced downtime, ultimately improving the overall data management processes. By automating fault recovery using AI, organizations can proactively address issues and minimize the impact of failures, enabling smoother and more reliable data processing operations. This research has the potential to provide valuable insights and best practices for designing resilient data pipelines, ultimately contributing to enhanced data management practices in organizations.

This thesis explores the designing of fault-tolerant data pipelines in Chapter 10, discusses the implementation of fault tolerance in Data Pipelines in Chapter 11, and automation of fault recovery using AI in Chapter 12. For this objective, the following research questions and sub-questions are studied in the included publications:

- **RQ3:** How to design and implement fault-tolerant data pipelines?
  - **RQ3.1.** What are the components required for data pipelines?
  - **RQ3.2.** How to implement fault tolerance in data pipelines?

These objectives collectively guide the research process and structure the subsequent chapters of the thesis. Through systematic exploration and analysis, this study aims to fulfill each objective, thereby addressing the broader research aim comprehensively. Fig. 3.2 shows the RQs and sub-RQs discussed in the thesis.



RQ1. What are the data management challenges, limitations of existing solutions, and open research problems in the field?

RQ 1.1. What are the data management challenges?
RQ 1.2. What are the limitations of the existing solutions for data management challenges?
RQ 1.3. What remain as open challenges in data management?

RQ2. How can DataOps and Data pipelines alleviate the data management challenges?

RQ 2.1. How DataOps contribute to data management?
RQ 2.2. How data pipelines contribute to data management?
RQ 2.3. What is the role of bidirectional data pipelines in data management?

RQ3. How to design and implement a fault-tolerant data pipelines?

RQ 3.1. What are the components required for a data pipelines?
RQ 3.2. How to implement fault-tolerance in data pipelines?

Figure 3.2: Research Questions

To achieve these objectives presented in the previous section, this thesis utilizes a range of different research methods, such as systematic literature reviews, multi-vocal literature reviews, multiple case studies, and empirical evaluations in collaboration with multiple companies. In the next sections, we provide an overview of these methods and their collaborations with the industry.

3.2 Research Context

The research conducted for this thesis was carried out in collaboration with Software Center [1], an initiative promoting research projects in close partnership with both industry and academia. The vision of the Software Center is to accelerate the digitalization of the European software-intensive industry. This industry collaboration focuses on advancing practices beyond traditional agile methods to encompass DevOps, A/B experimentation, and the integration of artificial intelligence in software development. With a focus on making use of

[1] https://www.example.com/software-center

the resources, expertise, and collaborative networks facilitated by the Software Center, all three objectives of this thesis were positioned within its framework. It is noteworthy that all resulting publications from this research have involved collaboration with companies affiliated with Software Center. This highlights the critical role of industry-academia collaboration in shaping the research agenda, methodology, and outcomes to address real-world challenges. By collaborating with 17 companies and 5 universities as strategic partners, Software Center provides a platform for collaboration and knowledge exchange. The Software Center operates with a sprint model that allows for frequent validation opportunities. In these sprints, teams work on short, focused development cycles to deliver incremental improvements and innovations. This model enables real collaboration among stakeholders, including researchers, industry partners, and other contributors, to drive continuous learning and improvement. Through this unique collaboration, participants engage in hands-on activities, share knowledge, and work closely together to address industry challenges and drive digital innovation. Selected embedded system companies specializing in Artificial Intelligence, particularly machine learning/deep learning, participated in the research, based on their domain expertise and maturity in AI adoption. To maintain confidentiality, participating companies remain anonymous due to the nature of the research, which often addresses technology limitations and development pitfalls.

In our research methodology, we engaged both primary and secondary case companies to ensure a comprehensive understanding of the data management challenges and practices in AI-enhanced embedded systems.

Primary case companies refer to those directly involved in the research collaboration, actively participating in data collection, validation, and solution development processes. These companies are typically industry partners collaborating with researchers to address specific challenges. Their involvement is deep and hands-on, providing firsthand insights, access to real-world data, and validation opportunities. The primary case companies play a crucial role in shaping the research agenda, guiding the direction of the study, and co-creating solutions.

Secondary case companies, on the other hand, are not directly involved in the research collaboration but serve as valuable sources of information and comparison. These companies may have similar characteristics or face similar challenges to the primary case companies but are not actively engaged in the research process. Instead, their data, practices, or experiences are analyzed and compared with those of the primary case companies to enrich the research findings and provide broader insights into the industry landscape.

The distinction between primary and secondary cases aligns with established research methods, such as case study research. Case study methodology emphasizes the in-depth investigation of real-life phenomena within their context, with primary cases representing the main focus of inquiry and secondary cases providing additional context and comparative analysis. By employing both primary and secondary cases in our research, we ensure a comprehensive understanding of data management challenges and practices in AI-enhanced embedded systems, enriched by diverse perspectives and insights from multiple sources.

3.2.1 Primary case companies

Companies A, B, C, D, E, and F are marked as primary case companies as they actively participated in this research by allowing collaboration through interviews, workshops, interactive sessions, weekly meetings, and action research.

Company A is a developer of an artificial intelligence platform designed to make the production of commercially viable AI applications swift, methodical, and scalable. The company's platform enables their clients ranging from startups to large-scale enterprises to pursue the benefits of integrating AI into their systems.

Company B is a multinational company within the telecommunication industry that distributes easy-to-use, adaptable, and scalable services that enable connectivity.

Company C is from the automobile domain manufacturing their cars and does analytics based on the data from multiple manufacturing units, delivery units, and repair centers for identifying poor-performing models.

Company D focuses on automotive engineering and depends on Company C which does modular development, advanced virtual engineering, and software development for them.

Company E is within the manufacturing domain having more than 19,000 employees and they manufacture and market pumps. They have standards in terms of innovation, efficiency, reliability, and sustainability.

Company F is a manufacturer of network-based solutions in the areas of physical security and video surveillance. The company is active in many market segments, including transport, infrastructure, trade, banking, education, state and municipality, and industry.

3.2.2 Secondary case companies

Companies G to M also contributed to the research through cross-company workshops. The reflections from the informants from these companies have helped in confirming the identified challenges and validity of the solution.

Company G works as a sales engagement platform that primarily enables and optimizes communication between sales representatives and potential prospects. Sales communication occurs in natural language via different communication channels, including emails.

Company H is a multinational technology company that develops, manufactures, licenses supports, and sells computer software, personal computers, consumer electronics, and services.

The company I is a global software company that develops both software and hardware solutions for home consumers.

Company J is a multinational automotive manufacturer and supplier of transport solutions. As the company's products are continuously growing in complexity and software size, the company is looking for strategies to prioritize its R&D effort and deliver more value to its customers.

Company K is a global car manufacturer that uses AI for building autonomous drive technology.

Company L is a global automotive manufacturer that collects and analyzes large amounts of data from vehicles and hundreds of thousands of connected vehicles to develop increasingly more intelligent computer models that can identify patterns hidden from human view and capabilities.

Company M is a manufacturer of power tools, industrial and construction technology, and packaging technology. They apply big data and machine learning to their products and services to create AI solutions that are safe, robust, and explainable.

Company N is a multinational packaging industry that manufactures machines and materials for disposable packaging for milk, juice, and other liquid foods.

3.3 Research Approach

Qualitative research methodologies offer a robust and comprehensive approach to understanding the multifaceted challenges and opportunities encountered by both practitioners and organizations [47]. It enables researchers to capture the diverse perspectives, needs, and priorities of stakeholders involved in data management initiatives. By conducting semi-structured interviews and collaborative workshops with industry practitioners, we investigated stakeholders' perceptions, challenges, and desires related to data management, facilitating the identification of solutions that align with stakeholders' interests and objectives [48]. Further, it allows for a deep exploration of the organizational contexts within which data management practices operate and allows us to gain insights into the organizational structures, cultures, policies, and practices that shape data management strategies and decisions.

Empirical evidence corroborates the utility of qualitative research in identifying best practices, emerging trends, and challenges in data management within industry settings [49]. Through qualitative data analysis techniques such as thematic coding and content analysis, researchers can derive insights from industry experiences, highlighting successful strategies, common problems, and areas for improvement in data management practices. Further, it fosters collaboration between researchers and industry partners in addressing real-world data management challenges [50]. By engaging company stakeholders as active participants in the research process, researchers can design models, tools, or frameworks that are tailored to industry needs and preferences, ensuring the relevance and applicability of research outcomes in practical settings.

In summary, qualitative research allows us to explore organizational contexts, understand stakeholder perspectives, identify best practices and challenges, design and implement solutions, and contribute to organizational success in the dynamic and evolving field of software engineering.

The principal advantage of using qualitative research methods is that they force the researcher to delve into the complexity of the problem rather than abstract it away [49]. Empirical data is the information that is collected utilizing the senses, particularly by observation and documentation of patterns and behavior through experimentation to answer the research question [49]. Both data collection and data analysis can be qualitative as well as quantitative [51]. The qualitative data collection process entails the generation of massive amounts of data [52]. The audio- or video-recording data collection method is followed by the transcription before the data analysis [51].

3.3.1 Case Study

A case study serves as a comprehensive examination of a specific instance or a limited number of instances of a phenomenon, delving deep into the intricacies of the subject [53]. The primary objective of conducting a case study is to analyze contemporary phenomena within their authentic real-world setting, particularly when the boundaries between the phenomenon and its context are blurred or intricate [54]. This research method offers several advantages, including the ability to collect and analyze data within the context of the phenomenon, the integration of both qualitative and quantitative data for analysis, and the capacity to capture the complexities of real-life situations, enabling a more profound investigation of the phenomenon.

Despite its strengths, case studies also come with certain limitations, such as potential issues related to rigor, challenges in data analysis, and limitations in making generalizations based on findings and conclusions [55]. Research strategies within case studies are typically categorized as exploratory, descriptive, explanatory, and improvement-oriented, depending on the research purpose. While case studies were initially employed for exploratory purposes, they have since been utilized for descriptive, improvement-focused, and explanatory research objectives as well [56].

In addressing exploratory research questions, the case study approach proves to be particularly suitable. Similarly, for descriptive research inquiries, a case study may be a viable option if sacrificing the representativeness of a sampling-based study can lead to a more realistic depiction of the subject. On the other hand, when representativeness is paramount, a survey method might be a more appropriate choice. Explanatory research questions can also be explored through case studies, although the evidence provided is not a statistically significant quantitative analysis of a representative sample, but rather a qualitative understanding of how phenomena operate within their specific context. For research purposes focused on improvement, the action research strategy emerges as a natural choice, often considered a variant of case study research [56].

In our study, we opted for an exploratory case study to uncover the challenges associated with data management practices in a real-world company setting. This approach allowed us to intricately capture the complexities involved in data management within the context of an Embedded system company scenario, providing valuable insights for our research endeavor.

3.3.2 Action Research

Action research can be seen as one alternative to intensify the conducting of important experimental studies with results of great value while investigating Software Engineering practices in depth [57]. It focuses on making direct interventions or actions in real-world settings to understand and improve practices. The Action Research approach typically means that researchers engage with a company over time and during a process. Problem owners are an inevitable part of action research since they share their skills, domain knowledge, and experiences [58] [59]. The main objective of action research in software engineering is to simultaneously solve a real-world problem and explore the experiences and results of problem-solving [60]. We chose the action research method for this study as the participatory aspect of it allowed us to systematically determine, and define the problem with data management practices, and make a solution proposal in the context of an investigation. Moreover, it allowed us to actively participate in further steps of applying the solution in real-time, which is termed as action [58] [59]. The action research process cycle consists of five stages namely (1) diagnosis, (2) action planning and designing, (3) action taking, (4) evaluation, and (5) specifying learning [58] [59]. Action research is advantageous as it has the potential to deliver robust and practical knowledge to a wide community of management and organization scholars [61].
Some key characteristics of action research are:

- **Collaboration:** Action research involves collaboration between researchers and practitioners in the host organization. This collaborative approach helps in addressing real-world problems and implementing practical solutions.

- **Iterative Process:** Action research is a cyclic process where researchers diagnose a problem, plan and implement interventions, collect and analyze data, and reflect on the outcomes. This iterative nature allows for continuous improvement and learning.

- **Participatory Approach:** Action research emphasizes the active participation of all stakeholders, including researchers, practitioners, and organizational members. This participatory approach ensures that the research is relevant and meaningful to the host organization.

- **Focus on Improvement:** The primary goal of action research is to improve existing practices, processes, and outcomes within the host organization. By implementing interventions and studying their effects, action researchers aim to bring about positive change.

- **Flexibility:** Action research offers flexibility in design and implementation, allowing researchers to adapt their approach based on feedback and changing circumstances. This adaptability is crucial in complex and dynamic organizational environments.

The action research provides a systematic approach to addressing complex problems, generating actionable insights, and fostering collaboration between academia and industry [62]. By involving stakeholders in the research process, action research helps build trust, promote learning, and drive sustainable change within organizations [62]. Additionally, action research can lead to the development of practical solutions, tools, and methods that have a direct impact on improving processes and practices.

We used a combination of case studies and action research as it offered a powerful methodological approach for investigating the challenges, solutions, and implementation strategies related to data management, DataOps, and data pipelines for AI-enhanced embedded systems. With this approach, we not only generated new knowledge and insights but also translated them into actionable recommendations that drove positive change within organizations. Action research and case studies can complement each other in research projects. Action research benefits from the detailed insights and rich data provided by case studies, while case studies benefit from the practical interventions and iterative approach of action research [63]. By combining elements of both methodologies, researchers can gain a deeper understanding of real-world phenomena [64]. Action research can inform the design and focus of case studies, while case studies can provide valuable input for the interventions and actions taken in action research projects [62]. Using both action research and case studies in a research project can lead to a more comprehensive and holistic understanding of complex issues in software engineering. The synergy between the two methodologies helps researchers address research questions from multiple perspectives and generate valuable insights for both academia and industry.

In summary, while action research emphasizes intervention and change within organizations, case studies offer detailed and contextualized analyses of specific cases. The synergy between these two methodologies can enhance research outcomes and contribute to a more comprehensive understanding of software engineering practices and challenges.

3.4 Research Techniques

The research techniques are used to gather empirical data necessary to analyze the actions in real-world industrial settings [65]. Research techniques such as semi-structured interviews, and literature reviews are appropriate for practical situations in which a fuller understanding of behavior, the meanings and contexts of events, and the influence of values on choices are useful for researchers.

3.4.1 Systematic literature review

A systematic literature review (SLR) is a rigorous and methodical approach in software engineering research that plays a crucial role in synthesizing existing literature to identify solutions to specific problems, such as data management.

SLRs ensure the reliability and validity of review findings, reducing bias and providing credible insights into current research trends and gaps in the literature by using predefined search criteria, systematic screening processes, and rigorous data extraction methods [66]. The systematic nature of SLRs allows researchers to comprehensively analyze a wide range of studies, identify common themes, and draw meaningful conclusions that can inform both academic research and industry practices. Furthermore, conducting an SLR helps researchers establish a strong foundation of knowledge on a particular topic, enabling them to build upon existing research and contribute new insights to the field of software engineering. By synthesizing diverse sources of information, SLRs enable researchers to identify best practices, emerging trends, and areas where further research is needed, thus guiding the development of innovative solutions and approaches in software engineering [67]. The findings of an SLR not only contribute to academic discourse but also have practical implications for software engineers, project managers, and other stakeholders in the software development process, empowering them to make informed decisions based on evidence-based insights. In this way, systematic literature reviews serve as a cornerstone of evidence-based practice in software engineering, driving advancements in the field and facilitating informed decision-making at both the research and practical levels.

3.4.2 Interviews

In interview-based data collection, the researcher asks a series of questions to a set of subjects about the areas of interest in the case study. Data collection through interviews is important in case studies [68]. The dialogue between the researcher and the subject(s) is guided by a set of interview questions. The interview questions are based on the topic of interest in the case study. That is, the interview questions are based on the formulated research question. The questions can be asked either to a group (focus group interviews) or to individual practitioners. Questions that allow and invite a broad range of answers and issues from the interviewed subject are called open-ended, while closed offer a limited set of alternative answers. Interviews can be divided into unstructured, semi-structured, and fully structured interviews [69]. In an unstructured interview, the interview questions are formulated as general concerns and interests of the researcher. In this case, the interview conversation will develop based on the interest of the subject and the researcher, whereas in a fully structured interview, all questions are planned and all questions are asked in the same order as in the plan. In many ways, a fully structured interview is similar to a questionnaire-based survey. In a semi-structured interview, questions are planned, but they are not necessarily asked in the same order as they are listed. We chose semi-structured interviews as they are helpful in the means of data collection because of two primary considerations. First, they are well suited for the exploration of the perceptions and opinions of respondents regarding data management issues and enable probing for more information and clarification of answers. Second, the opportunities for face-to-face contact with a researcher stimulate interest in the project and establish a sense of rapport between respondents and the researchers [70].

3.4.3 Observation

Observation is the conscious noticing and detailed examination of participants' behavior in a naturalistic setting [71]. Observations can be conducted to investigate how a certain task is conducted by practitioners. There are many different approaches to observation. One approach is to monitor a group of practitioners with a video recorder and later on, analyze the recording, for example, through protocol analysis [72] [73]. Another alternative is to apply a "think aloud" protocol, where the researcher is repeatedly asking questions like "What is your strategy?" and "What are you thinking?" to remind the subjects to think aloud. Observations in meetings are another type, where meeting attendants interact with each other and thus generate information about the studied object. An alternative approach is where a tool for sampling is used to obtain data and feedback from the participants [74]. While experiencing what is going on in a research site, researchers need to observe this and make detailed notes, called field notes, about the people, the concepts they discuss, and the interactions that occur [71]. Participant observation was performed, and field notes were taken during the action research. Observation as a data collection method can be structured or unstructured. In structured or systematic observation, data collection is conducted using specific variables and according to a pre-defined schedule. Unstructured observation, on the other hand, is conducted in an open and free manner in the sense that there would be no pre-determined variables or objectives [56]. The unstructured observation was used in this research as the observation mainly happened during the weekly stand-up meetings, pair programming, and weekly presentation of results.

3.4.4 Multi-vocal literature review

The multi-vocal literature review is used to explore and summarize existing evidence concerning a particular topic [67] [75] [76] and to identify gaps and limitations of existing practices. A Multivocal Literature Review (MLR) is a form of a Systematic Literature Review (SLR) [77] which includes the grey literature (e.g., blog posts, videos, and white papers) in addition to the published (formal) literature (e.g., journal and conference papers) [75]. MLRs are useful for both researchers and practitioners, since they provide summaries of both the state-of-the-art and –practice in a given area. Grey literature by the practitioners was ignored tagging them as "unscientific" while practitioner interviews are done and reported by researchers have, for long, been considered as academic evidence in empirical software engineering. MLR is developed to lift such a double standard by allowing rigorously conducted analysis of practitioners' writings to enter the scientific literature [75].

We employed systematic literature review, semi-structured interviews, observation, and multi-vocal literature reviews as research techniques. For the first objective of this research, we did a systematic literature review to identify the solutions for the data management challenges we identified through semi-structured interviews. For the second objective, we wanted to collect empirical evidence about the challenges associated with existing data management practices from the practitioners. We chose semi-structured interviews as it allows informants the freedom to express their views on their terms. Moreover, semi-structured interviews allow us to gather in-depth, comparable, and reliable empirical data. One of the data management practices we identified was relatively new, and there was not much peer-reviewed literature that discussed it. Therefore, we chose a multi-vocal literature review to frame a definition for that particular data management practice. We used unstructured observation as a research technique, as we were allowed to attend the weekly team meetings and other discussions. Thus, notes were taken during the weekly stand-up meetings, pair programming, and weekly presentation of results.

3.5 Data Analysis

Qualitative research yields mainly unstructured text-based data. These textual data could be interview transcripts, observation notes, diary entries, or medical records. In some cases, qualitative data can also include a pictorial display, audio or video clips (e.g. audio and visual recordings of patients, radiology film, and surgery videos), or other multimedia materials. Therefore, the data analysis methods should be a dynamic, intuitive, and creative process of inductive reasoning, thinking, and theorizing.

3.5.1 Qualitative Data Analysis

Data analysis in qualitative research is defined as the process of systematically searching and arranging the interview transcripts, observation notes, or other non-textual materials that the researcher accumulates to increase the understanding of the phenomenon [78]. The process of analyzing qualitative data predominantly involves coding or categorizing the data. Coding merely involves subdividing a huge amount of raw information or data and subsequently assigning them into categories [79]. Thematic coding using the NVivo tool and open coding are the two types of coding used in this licentiate thesis. Thematic coding is a type of qualitative data analysis that finds themes in the text by analyzing the meaning of words and sentence structure. As NVivo is a thematic analysis software that helps you automate the data coding process, there was no need to set up themes or categories in advance [80]. Open coding is a manual coding technique that starts from scratch and creates codes based on the qualitative data itself. Codes are manually created in such a way that it covers the entire transcript. These codes are then applied to the remaining transcripts and necessary adjustments are made so that the codes apply to all transcripts in the study [81].

3.6 Threats to Validity

This section discusses threats to validity regarding how our research questions were answered.

3.6.1 Construct Validity

Construct Validity includes two components: the measure should be exhaustive, and the measure should be selective in that it only covers aspects of the target theoretical construct. To ensure construct validity, a few cases were excluded from the results, as some interviewers did not have a proper understanding of the discussed concepts. As a result of the screening process, our study has some limitations with several interviews. However, this limitation can be counted as an opportunity for further inquiry in future works. To reduce researcher bias, the interviews were conducted by a minimum of two researchers. Further, before the interviews, we developed the semi-structured interview guide and distributed it among the interviewees. A short description of the topic to explore is sent to the interviewees before the interview. During the interview, we again explained the topic of the study as an introduction. We rephrased the question whenever the response became off-topic, or asked them to elaborate when we received ambiguous answers. Further, while analyzing the transcripts if there is any confusion or lack of clarity, we contacted the interviewees to resolve this problem.

3.6.2 Internal Validity

Internal validity is defined as the degree to which the observed outcome represents the truth in the population we are studying and, thus, is not due to methodological errors [58]. The results of this thesis could potentially be affected by this threat since the results and strategies associated with RQ2 and RQ3 were developed in the company context. As the researcher only had limited access to the descriptions of the strategies, it is not possible to investigate if other factors were more influential to the final result than the proposed strategies. To minimize internal validity threats, one of the co-authors, who has in-depth knowledge about the data processed in the company, was asked to validate the findings. Further, the findings were validated through the steering committee at the respective companies.

3.6.3 External Validity

The presented work is derived from the cases studied with different teams in the domains of manufacturing, automobile, and telecommunication. Some parts of the work can be seen in parts of the company differently. All the terminologies used in the companies are normalized and the implementation details are explained with the necessary level of abstraction [82]. We do not claim that the opportunities and challenges will be the same for industries from different disciplines.

Chapter 4

Contributions of this thesis

This chapter describes how the included and the related publications are connected and contribute to a broader understanding of data management in AI-enhanced Embedded Systems. First, we provide a general overview of the research projects conducted in this doctoral study (that are connected to the objectives of this thesis). Second, we provide a summary of the study, the research method, and the main results of each included publication. Finally, we provide a summary of the related publications that are not included in this thesis. Fig 4.1 shows the relationship between the included papers (A, B, C, D, E, F, G, H, I, and J) and the objectives/research questions.

4.1 General Overview

Objective 1: Identify the data management challenges

The first objective was initially investigated from the academic perspective, with literature reviews and evaluations based on the results in papers **K** and **L**. These first results led us to a case study in collaboration with six industry partners. By systematically analyzing the data management process throughout the data lifecycle in the context of deep learning, we identified challenges at each stage. Mapping these challenges to specific data life-cycle stages provided a structured understanding of the complexities involved. We conducted a comprehensive review of existing literature on data management to identify solutions. By synthesizing and analyzing these solutions, we gained insights into effective strategies and approaches for addressing data management challenges. Then we examined existing solutions from other domains and compared them with the unique requirements and characteristics of deep learning tasks to analyze why these solutions are not sufficient to address the data management challenges specific to deep learning. We classified data management challenges based on the availability of solutions to provide insights into the current state of research and practice in the field. The study is detailed in paper **A**. This classification can be used for prioritization and allocation of resources towards addressing the most significant challenges where solutions are inadequate.

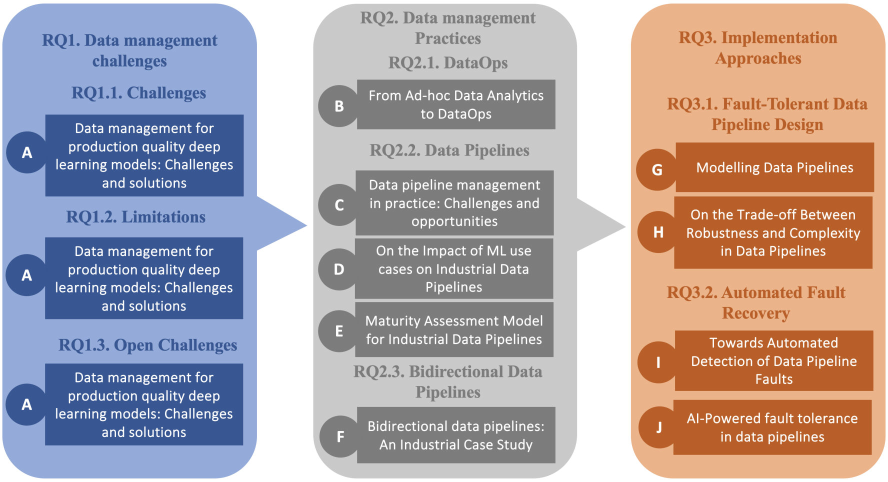

RQ1. Data management challenges
RQ1.1. Challenges
A
Data management for production quality deep learning models: Challenges and solutions
RQ1.2. Limitations
A
Data management for production quality deep learning models: Challenges and solutions
RQ1.3. Open Challenges
A
Data management for production quality deep learning models: Challenges and solutions
RQ2. Data management Practices
RQ2.1. DataOps
B
From Ad-hoc Data Analytics to DataOps
RQ2.2. Data Pipelines
C
Data pipeline management in practice: Challenges and opportunities
D
On the Impact of ML use cases on Industrial Data Pipelines
E
Maturity Assessment Model for Industrial Data Pipelines
RQ2.3. Bidirectional Data Pipelines
F
Bidirectional data pipelines: An Industrial Case Study
RQ3. Implementation Approaches
RQ3.1. Fault-Tolerant Data Pipeline Design
G
Modelling Data Pipelines
H
On the Trade-off Between Robustness and Complexity in Data Pipelines
RQ3.2. Automated Fault Recovery
I
Towards Automated Detection of Data Pipeline Faults
J
AI-Powered fault tolerance in data pipelines

Figure 4.1: Relationship between the papers and the research questions

Objective 2: Explore the data management practices that can alleviate data management challenges

While the first objective consisted of the identification of challenges, through cross-company workshops we validated whether these challenges were recognized by the experts from the companies in the Software Center, which are mainly in the embedded systems domain. Then we did a case study(Paper **B**) to identify the principles, methodologies, and tools associated with DataOps, which laid the foundation for exploring its applicability in alleviating data management challenges. We studied the evolution of data analytic teams' infrastructure and processes towards DataOps and identified trends, patterns, and best practices that contribute to more efficient and effective data management. We developed a stairway model to depict the stages of evolution towards DataOps to help companies assess their current maturity level and identify areas for improvement which serves as a roadmap for organizations to progress towards more advanced data management practices, overcoming challenges encountered at each stage.

From the results of the study in Paper **B**, we realized the importance of data pipelines and therefore conducted a follow-up case study with multiple companies and identified the opportunities of having a dedicated data pipeline as well as the key challenges associated with data pipeline management. This analysis provides a rationale for investing resources in optimizing data pipeline infrastructure to enhance data management capabilities. We developed a taxonomy of data pipeline challenges including infrastructural, organizational, and technical(Paper **C**) so that organizations can identify the root causes of data pipeline inefficiencies and develop targeted strategies for improvement. We also identified the determinants used to evaluate the maturity of data pipelines and designed a data pipeline maturity assessment model (Paper **D**) that serves as a diagnostic tool for identifying areas of strength and areas requiring improvement in data pipeline management practices. Further, we wanted to assess the role of ML (Paper **E**)in shaping modern data management practices for which we did action research and categorized data pipelines into four types based on the criticality of the application and purpose to provide a structured framework for understanding the diversity of data management practices. We also did a comprehensive analysis of the determinants shaping data management practices to identify key factors influencing the development and deployment of data pipelines.

During a cross-functional workshop, practitioners mentioned bidirectional data pipelines, so we did a study to identify the key differences between unidirectional and bidirectional data pipelines, and we also outlined a roadmap for smooth migration from unidirectional to bidirectional data pipelines. Comparing unidirectional and bidirectional data pipelines and highlighting their respective advantages and disadvantages (Paper **F**) is highly relevant to data management, as it helps organizations understand the trade-offs and benefits associated with each approach. We also discussed the significance of bidirectional data pipelines and their importance in facilitating real-time data exchange and enabling more dynamic data management processes.

Objective 3: Investigate the implementation approaches for improved data management

The Third objective draws on the insights from the second objective to develop more effective implementation approaches. By aligning with proven data management practices, this objective can ensure that its implementation approaches are practical, efficient, and aligned with industry standards. We conducted a case study to understand the challenges associated with data management using existing data pipelines and developed a conceptual model(Paper **G**) that offers a structured framework for building data pipelines, especially for applications like machine learning/deep learning models. This model incorporates automatic monitoring, fault detection, mitigation, and alarming techniques, enhancing data management by ensuring robustness and reliability. We validated the conceptual model through another case study with leading companies from various domains (manufacturing, telecommunication, and automobile) to add credibility and practical applicability to the proposed model. As a follow-up study, we did action research to improve the reliability of data management practices like data pipelines. For this, we identified typical faults in data pipelines and mitigation strategies adopted by practitioners for fault tolerance. By addressing common faults proactively, the impact of pipeline failures can be minimized, thus improving data management reliability. We also proposed a fault-tolerant data pipeline model (Paper **I**) capable of automatically detecting and mitigating common faults, contributing to improved data management. As an upgraded version of the above model in Paper **H**, we introduced an AI-powered 4-stage model for automated fault recovery in data pipelines (Paper **J**). We also validated the proposed fault recovery model using industrial datasets, demonstrating its effectiveness and applicability in real-world scenarios. It not only provides empirical evidence of the model's performance and reliability but also bolsters confidence in its adoption for improved data management.

4.2 Included publications

4.2.1 *Paper A: Data management for production quality deep learning models: Challenges and solutions*

4.2.1.1 Summary of the Paper

The paper discusses the challenges encountered in data management for deep learning models in real-world industrial settings. Through a multi-case study and a systematic literature review, the authors identified data management challenges across various phases of the data lifecycle. Challenges such as lack of labeled data, data granularity, shortage of diverse data samples, data sharing and tracking methods, and data storage compliance with GDPR were highlighted. The study also classified these challenges based on the data lifecycle phase in which they occur and provided insights into the implications and empirical basis of each challenge.

4.2.1.2 Research Method

The research employed a three-step approach consisting of an interpretative multi-case study, a systematic literature review, and a validation study. The interpretive multi-case study involved interviews with experts working on deep learning systems across different domains to identify data management challenges. A systematic literature review was then conducted to identify potential solutions for the challenges. Finally, a validation study was conducted with the same experts to validate the solutions identified in the literature.

4.2.1.3 Main Results

The main results of the study include the identification of key data management challenges faced by practitioners developing deep learning systems. These challenges were categorized based on the data lifecycle phase in which they occur, providing a comprehensive understanding of the issues encountered. The study also highlighted the solutions proposed in the literature for addressing these challenges, along with their limitations and applicability in practical settings. Overall, the research contributes to the ongoing discussion on data management challenges in deep learning models and provides insights for future research directions in this field.

4.2.2 *Paper B: From Ad-Hoc Data Analytics to DataOps*

4.2.2.1 Summary of the Paper

The paper discusses the evolution of data analytics processes from ad-hoc methods to DataOps, which aims to automate and optimize data collection, validation, and verification processes. It provides insights from a case study at Ericsson on how multiple data analytic teams evolved their infrastructure and processes towards DataOps. The paper also presents a stairway model showing the different stages of evolution in data strategy.

4.2.2.2 Research Method

The research methodology adopted for the study includes a Multivocal Literature Review to gather insights on DataOps, interviews with experts in the field, and an interpretive single-case study at Ericsson. The study focused on defining DataOps, analyzing the evolution of data analytic processes, and identifying challenges and requirements at each stage.

4.2.2.3 Main Results

DataOps is defined as an approach that accelerates the delivery of high-quality results by automating and orchestrating the data lifecycle stages. It incorporates best practices from Agile software engineering and DevOps for efficient analytics governance. Further, the study identified five stages in the evolution of data strategy, from ad-hoc data analysis to fully implemented DataOps. Each stage represents a progression in data collection, processing, and automation towards optimizing the end-to-end data analytic lifecycle. Furthermore, the research highlighted challenges such as organizational restructuring, lack of skilled teams, and data silos in implementing DataOps. The study emphasized the need for continuous testing, monitoring, automation, collaboration, and orchestration in the data analytics process to achieve the goals of DataOps.

4.2.3 *Paper C: Data Pipeline Management in Practice: Challenges and Opportunities*

4.2.3.1 Summary of the Paper

The paper explores the challenges and opportunities of implementing and managing data pipelines in real-world settings. It discusses the importance of data pipelines in data-driven organizations, the challenges faced in their implementation, and the benefits they bring in terms of automating data-related activities. The research includes a qualitative multiple-case study with three companies in the telecommunication and automobile domains to identify key challenges and benefits associated with data pipeline management.

4.2.3.2 Research Method

The research methodology adopted for conducting the study involved a qualitative approach through a case study. Data was collected through interviews and meetings with representatives from the companies. The study aimed to answer the research question on practical opportunities and challenges associated with the implementation and maintenance of data pipelines at the industry level.

4.2.3.3 Main Results

The main results of the study included the identification of challenges in data pipeline management, such as data quality issues, infrastructure challenges, and organizational barriers. The study also highlighted the benefits of data pipelines in terms of data accessibility, time and effort savings, improved traceability, and standardized data workflow. The research emphasized the importance of data pipelines in supporting DataOps culture within organizations. Challenges and opportunities in integrating new data sources, scalability, infrastructure complexity, data quality, and operational errors in data pipelines were discussed.

4.2.4 *Paper D: Impact of ML Use Cases on Industrial Data Pipelines*

4.2.4.1 Summary of the Paper

The paper explores the impact of Machine Learning (ML) use cases on industrial data pipelines by analyzing six data pipelines from different companies. It addresses the importance of data quality, data preprocessing, data storage requirements, data pipeline elements, performance efficiency, and continuous monitoring for both ML-influenced and non-ML data pipelines. The study categorizes the use cases based on criticality and purpose, identifying ML use cases as having a significant impact on data pipelines.

4.2.4.2 Research Method

The research methodology employed a multiple-case study approach involving interviews, observation, and document analysis to study the six data pipelines. Data were collected through semi-structured interviews and weekly meetings with company representatives. The data collected was analyzed, categorized, and validated through follow-up meetings and feedback from diverse teams within the companies. Guidelines for conducting interviews and focus groups in a virtual setting were followed to ensure the validity and reliability of the study.

4.2.4.3 Main Results

The main results of the study indicate that ML use cases and high-criticality non-ML use cases demand more sophisticated data pipelines compared to low-criticality non-ML use cases. Determinants such as big data requirements, data preprocessing efforts, data quality standards, data storage needs, data pipeline elements, performance efficiency, and continuous monitoring play crucial roles in assessing the impact of ML use cases on data pipelines. The study emphasizes the significance of continuous monitoring, fault detection, and mitigation for ensuring the smooth operation of data pipelines serving ML models.

4.2.5 *Paper E: Maturity Assessment Model for Industrial Data Pipelines*

4.2.5.1 Summary of the paper

The paper discusses the importance of data pipelines in data-driven organizations and the challenges faced in assessing and enhancing the maturity of data pipelines. It introduces a maturity assessment model for evaluating the maturity of data pipelines in a staged manner from level 1 to level 5. The research focuses on developing the maturity assessment model based on five determinants and aims to help organizations assess their current data pipeline maturity, identify challenges, and provide recommendations for improvement.

4.2.5.2 Research Method

The research methodology involved conducting multiple qualitative case studies with organizational units of industrial enterprises. Semi-structured interviews were conducted with practitioners working on data pipelines to identify determinants for assessing the maturity of data pipelines. The research followed a structured approach to develop the maturity assessment model based on empirical data from the case studies.

4.2.5.3 Main results

The main results of the research include the development of a staged maturity assessment model for data pipelines based on five determinants: security, scalability, resiliency, robustness, and dependability. The findings show the evolution stages of data pipelines from level 1 to level 5, with each level representing a different maturity stage. Practical challenges and recommendations for improving data pipeline maturity were also identified based on the case study findings.

4.2.6 *Paper F: Bidirectional Data Pipelines: An Industrial Case Study*

4.2.6.1 Summary of the paper

The paper delves into the role and importance of bidirectional data pipelines in modern data-driven environments, focusing on the differences between unidirectional and bidirectional data pipelines. Through a qualitative multiple-case study approach, the research explores the benefits, challenges, and considerations essential for transitioning from unidirectional to bidirectional data pipelines. The study highlights the significance of bidirectional data pipelines in enhancing data consistency, improving workflow efficiency, facilitating real-time data synchronization, and facilitating seamless integration of disparate systems.

4.2.6.2 Research Method

The research methodology employed a qualitative approach centered around a multiple-interpretive case study with professionals from a multinational telecommunications vendor. Data collection involved semi-structured interviews, analysis of meeting notes, documentation, and presentations. Thematic coding using NVivo was used for data analysis, following a six-phase thematic analysis process. The study design and data collection were guided by software engineering case study guidelines to ensure a systematic and comprehensive exploration of bidirectional data pipelines within industrial contexts.

4.2.6.3 Main Results

The study identified key distinctions between two unidirectional data pipelines without a shared data transmission channel and bidirectional data pipelines. It highlighted challenges such as conflict resolution, data consistency, latency, and security concerns associated with bidirectional data pipelines. The paper emphasized the significance of bidirectional data pipelines in enhancing data consistency, improving workflow efficiency, facilitating real-time data synchronization, and enabling seamless integration of disparate systems. Additionally, the study underscored the role of bidirectional data pipelines in accelerating decision-making, facilitating data-driven strategies, and ensuring real-time data synchronization for informed decision-making in organizations.

4.2.7 *Paper G: Modelling Data Pipelines*

4.2.7.1 Summary of the paper

The paper discusses the importance and challenges of data pipelines in managing and processing data efficiently. It highlights the need for automation and fault detection in data pipelines to address the complexities involved in handling high-quality data from various sources. The research focuses on proposing a conceptual model of an end-to-end data pipeline that can be used as a standard language for communication between different data teams, enabling automation of monitoring and mitigation processes.

4.2.7.2 Research Method

The study adopts an exploratory case study approach to understand the challenges faced by practitioners in managing data and existing data pipelines. Qualitative data is collected through interviews and meetings with industry professionals from multiple companies in different domains. The research methodology involves formulating research questions, developing a conceptual model, and conducting validation studies internally within the telecommunication company and externally with two manufacturing companies. The study aims to validate the proposed conceptual model of the data pipeline through feedback and discussions with industry experts.

4.2.7.3 Main Results

The conceptual model of the data pipeline is validated through interviews with industry professionals, leading to agreements on the necessity of standard pipeline models and the automation of monitoring processes. The study identifies challenges in data management that can be partially or completely solved by implementing the proposed data pipeline model. While some challenges like data availability and data dependencies can be completely solved, others such as unreliable data pipelines and low storage capacity may require further enhancements. The research highlights the potential of fault-tolerant, automated, and traceable data pipelines in addressing data management challenges faced by data-driven companies.

4.2.8 *Paper H: On the Trade-off between Robustness and Complexity in Data Pipelines*

4.2.8.1 Summary of the Paper

The paper discusses the trade-off between the robustness and complexity of data pipelines in the context of data management processes such as data analytics and machine learning. It highlights the importance of ensuring robust data pipelines to maintain data quality and reliability. The study identifies essential components for robust data pipelines and analyzes the balance between robustness and complexity in order to optimize data pipeline performance.

4.2.8.2 Research Methodology

The study utilizes a multi-case study approach with an interpretive methodology to explore and understand real-world cases of data pipelines in various organizations. Data collection methods include semi-structured interviews and weekly meetings with experts from three case companies. The data analysis involves identifying stages of data pipelines, main purposes, similarities, and differences between use cases, and developing themes from interview transcripts to understand common components.

4.2.8.3 Main Results

The research findings demonstrate the crucial need for robust data pipelines to ensure high-quality data products in data-driven organizations. The study introduces a conceptual model for robust data pipelines, emphasizing the inclusion of connector capabilities such as fault detection, mitigation strategies, and authentication mechanisms. It highlights the challenges of balancing robustness and complexity in data pipelines and emphasizes the importance of prioritizing robustness to maintain data quality and reliability.

4.2.9 *Paper I: Towards Automated Detection of Data Pipeline Faults*

4.2.9.1 Summary of the paper

The paper discusses the importance of maintaining fault-tolerant and self-healing data pipelines in the context of large software-intensive organizations. It emphasizes the need for automated fault detection mechanisms and mitigation strategies to ensure the smooth flow of data and reduce the impact of faults at various stages of the data pipeline. The study explores real-world data pipelines in two companies, identifies common faults at different steps, and proposes corresponding mitigation strategies to improve data pipeline reliability and performance.

4.2.9.2 Research Method

The study employs an action research approach to investigate four data pipelines used in two companies. Through close collaboration with practitioners, the researchers explore the faults encountered during the development and maintenance of data pipelines. Conceptual modeling of fault-tolerant data pipelines is presented, focusing on automated fault detection and mitigation strategies at different stages. The researchers conduct weekly meetings, workshops, and discussions with data scientists, software developers, and other stakeholders to gather insights and feedback on the fault detection and mitigation mechanisms.

4.2.9.3 Main Results

The study identifies common faults such as data source failure, incompatible ingestion methods, unexpected data, changes in data formats, and human errors at various stages of the data pipeline. Mitigation strategies include sending alarms, data validation, defining standard schemas, lossless data transformation, and collaboration with subject-matter experts for accurate data interpretations. The implementation of fault detection components and mitigation strategies in a small slice of the data pipeline shows promising results in automating the recovery process and improving data pipeline resilience. This research contributes valuable insights for companies looking to enhance the reliability and fault tolerance of their data pipelines.

4.2.10 *Paper J: AI-Powered Fault Tolerance in Data Pipelines*

4.2.10.1 Summary of the Paper

The significance of preserving self-healing and fault-tolerant data pipelines in the context of AI-enhanced embedded systems is covered in the paper. It highlights how important automated fault detection systems and mitigation techniques are, particularly in settings where data is constantly changing. A four-stage model that can identify anomalies, pinpoint the fault, and suggest mitigation techniques is presented in the study. Through action research conducted at two companies, the 4-stage model is validated.

4.2.10.2 Research Method

To integrate AI-powered fault detection in data pipelines from two companies, the study uses an action research methodology. The researchers investigate the difficulties in locating errors in data pipelines through close collaboration with practitioners, proving the necessity of AI-powered procedures for fault detection and recovery. To get advice and comments on AI-powered fault recovery in data pipelines, the researchers held conferences, workshops, and talks with data scientists, software engineers, and other stakeholders.

4.2.10.3 Main Results

The AI-powered fault recovery in data pipelines is validated through interviews with industry professionals, leading to agreements on the necessity of implementing automated recovery of data pipelines. The study identifies challenges in detecting fault detection in the data pipelines using conventional methods, especially when the data is constantly evolving. While some familiar faults can be completely solved, other faults need special treatment. The research highlights the potential of fault-tolerant, automated, and traceable data pipelines in addressing data management challenges faced by data-driven companies.