Institute of Software Technology

University of Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Masterarbeit

Data Integration and Analysis Approaches in the Context of Automotive Events: A Case Study with Active Driver Assistance Systems

Markus Haug

Course of Study:	Softwaretechnik
Examiner:	Prof. Dr. Stefan Wagner
Supervisor:	Dr. Justus Bogner Dipl.-Ing. (FH) Matthias Gut Dipl.-Ing. (BA) Tobias Schreitmüller
Commenced:	2020-10-01
Completed:	2021-04-01

Abstract

Effective data analytics development is an important capability for modern enterprises. However, managing the data pipelines feeding these data analytics applications gives rise to new challenges, which are especially felt in the automotive domain.

This thesis investigates challenges for data pipeline management and suitable data analysis approaches for automotive event data in a case study with the Daimler Truck AG. A literature review identifies challenges for data pipeline management in general and in the automotive context. These challenges are compared to challenges identified through semi-structured interviews with data pipeline stakeholders at Daimler Truck AG. Approaches for handling the identified challenges are also discussed.

Furthermore, a prototype of a concrete data analysis on automotive event data is designed and developed. The prototype's performance is experimentally evaluated on an existing data set of automotive event data. This evaluation indicates a need for improvement of the approach. Based on the initial prototype, alternative analysis approaches are investigated and scope for future work is outlined.

Acknowledgements

I want to express my sincere gratitude to my supervisors **Dr. Justus Bogner**, **Matthias Gut**, and **Tobias Schreitmüller**. Without their guidance throughout the project, this thesis would not have been possible. Their expertise, insights, and editing shaped this thesis.

I also want to thank **Ingo Scherhaufer**, my team lead at Daimler Truck AG, who made this thesis possible in the first place.

Last but not least, I am eternally grateful for my parents' and brother's unwavering support and encouragement.

Contents

List of Figures

List of Tables

Acronyms

ABA	Active Brake Assist	17
ADA	Active Drive Assist	17
ADAS	Advanced Driver Assistance System	17
AEBS	Advanced Emergency Braking System	18
AMI	Adjusted Mutual Information	32
ARXML	AUTOSAR XML	54
ASGA	Active Sideguard Assist	17
AUTOSAR	Automotive Open System Architecture	54
BSIS	Blind Spot Information System	18
CTP	Common Telematics Platform	53
DBSCAN	Density-Based Spatial Clustering of Applications with Noise	41
DTW	Dynamic Time Warping	23
ECU	Electronic Control Unit	19
EDR	Event Data Recorder	18
GDPR	General Data Protection Regulation	19
GSR	General Safety Regulation	18
HAC	Hierarchical Agglomerative Clustering	58
IDL	Interface Description Language	54
JSON	JavaScript Object Notation	51
MDF	Measurement Data Format	54
OPTICS	Ordering Points To Identify the Clustering Structure	22
SAR	Situation Analysis Recorder	19
StVG	Straßenverkehrsgesetz	21
VIN	Vehicle Identification Number	20
VRDU	Video Radar Decision Unit	19

1 Introduction

This chapter motivates the need for the conducted research. It also states the problem, the research is trying to solve and describes the environment in which the case study was conducted. Finally, it outlines the structure of this thesis.

1.1 Motivation

For modern enterprises, effective development of data analytics has become an important capability [Kim20]. Usually, data collection and data integration [GHMT17] is handled by data pipelines, before the data is fed to data analytics applications. In addition to integrating data analytics into the data pipeline, practitioners face further challenges, such as compliance with data protection regulation [YLCT19]. Furthermore, large pre-labeled data sets are not always available for supervised data analytics, which may especially affect the automotive domain [HMD17; LKM+15]. Daimler Truck AG is one concrete example of this. It collects large amounts of event data from active driver assistance systems in its vehicles. Suitable data analysis methods need to be explored to utilize the full value of these event records.

1.2 Problem Statement

This thesis investigates the challenges that practitioners face when working with data integration and data analysis pipelines. Special attention is payed to challenges arising in data analysis efforts in automotive contexts.

Additionally, suitable data analysis approaches for automotive event data are studied. For this purpose, a concrete data analysis will be developed and integrated into an established data pipeline.

1.3 Case Study Environment

With around 100,000 employees and 378,500 vehicles sold in 2020, Daimler Truck AG is one of the largest commercial vehicle manufacturers worldwide. Daimler Truck AG is committed to improving road safety for all participants, a fact that is reflected in the wide range of driver assistance systems available in its vehicles. These driver assistance systems are developed by the driver assistance & brake systems department, TP/EMD. This thesis was conducted as a case study with the active safety group that is part of TP/EMD.

1.4 Structure

Chapter 2 elaborates on the background of this thesis. Chapter 3 highlights related studies into challenges for data pipeline management and automotive event data analysis. Chapter 4 describes the research process. In Chapter 5, the results obtained in this study are described. The evaluation of the developed prototype is covered by Chapter 6. The obtained results and the evaluation will be discussed in Chapter 7. Finally, Chapter 8 will conclude this thesis and highlight possible areas for future work.

2 Background

This chapter gives a short introduction to the technical background of event data recording in the context of ADAS. It also introduces some concepts of data pipelines and machine learning.

2.1 Driver Assistance Systems

Advanced Driver Assistance Systems (ADAS) are control systems embedded into vehicles that can assist drivers in performing various tasks related to driving a vehicle. Typically, these systems are able to influence vehicle speed and direction. For this purpose, ADAS can control the engine, the brakes, and the steering of the vehicle.

One ADAS available in commercial vehicles designed by Daimler Truck AG is Active Sideguard Assist (ASGA) [Dai20]. The system can assist drivers of those vehicles when turning by detecting pedestrians or cyclists in blind spots on the passenger side of the vehicle. Following a two stage warning concept, an optical indicator first informs the driver of the situation. This indicator is active as soon as any obstacle is detected in the monitored area. If the system further detects evidence of a driver's intent to perform a turn, the warning intensity is increased. The optical indicator changes color and flashes and an acoustical warning is played. In addition to the second stage warning, ASGA can engage the braking system to stop the vehicle. This brake can occur concurrently with the second stage warning or follow it.

Active Drive Assist (ADA) is another assistance system in commercial vehicles designed by Daimler Truck AG [Dai20]. The system supports partially autonomous driving. In longitudinal direction, the system can maintain a preset speed. If maintaining the speed is not possible, because there is another, slower vehicle in front ADA adapts vehicle speed in order to follow this other vehicle. The distance to the leading vehicle is based on a configured time gap. ADA can even stop the vehicle and start again. In lateral direction, the system can keep the vehicle on a configured position inside its lane.

The third driver assistance system available in commercial vehicles designed by Daimler Truck AG is Active Brake Assist (ABA) [Dai20]. This forward collision mitigation system aims to prevent crashes between the vehicle equipped with ABA and other vehicles or pedestrians in front of it. ABA also utilizes a multi-stage warning concept, similar to that of ASGA. Figure 2.1 shows that warning concept. The first stage consists of visual and acoustical signals which warn the driver about the danger of collision with a vehicle ahead. In the second stage, ABA brakes with about 50 % of the possible deceleration, called the haptic brake. In the third and final stage, the emergency braking, ABA brakes the vehicle using the full deceleration capability of the vehicle.

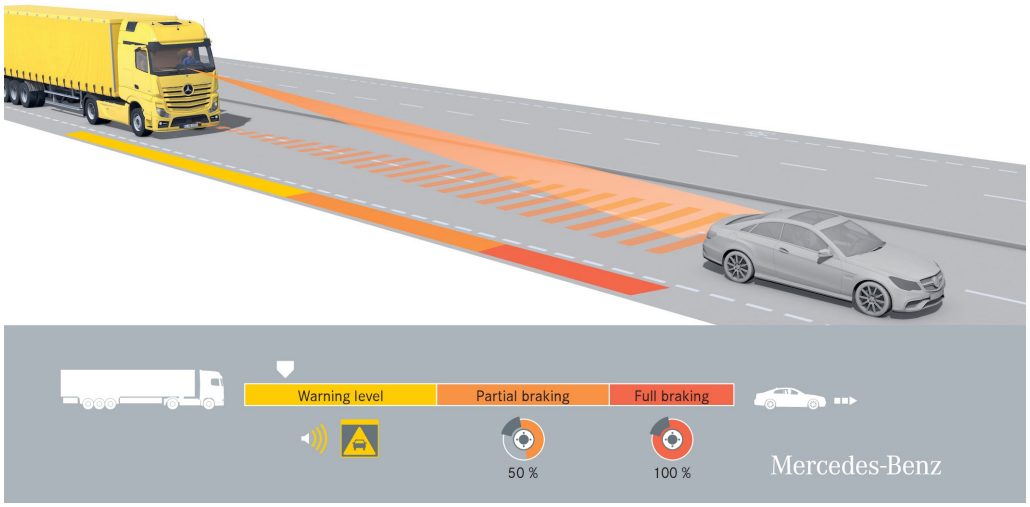


Figure 2.1: Multi-stage warning concept of ABA [adapted from Dai19]

As their name suggests, these assistance systems, however advanced they may be, can only assist the driver in controlling the vehicle. The driver stays in full control of the vehicle and may override the system at any time. However, that also means the responsibility for the vehicle rests solely with the driver.

In an attempt to eliminate accidents in road transport, the European Parliament and the Council of the European Union adopted EU regulation 2019/2144, the General Safety Regulation (GSR), in 2019. This regulations mandates that new vehicles must be equipped with multiple ADAS, such as Advanced Emergency Braking System (AEBS) and Blind Spot Information System (BSIS). Additionally, commercial vehicles will require an Event Data Recorders (EDRs) for type approval from January 7, 2026.

2.2 Event Data Recording

Event data recording devices in the broadest sense are devices that capture data about an event as it occurs. Event data recording technology has already been used in a variety of contexts, such as the black boxes used in aviation [CIM+01]. In this study, which deals with event data recording for ADAS in commercial vehicles, events usually correspond to an activation of an ADAS.

The recorded event data may be used for multiple purposes, such as reconstruction of an accident, defense against legal claims, or ongoing development.

2.2.1 Event Data Recording in Vehicles Designed by Daimler Truck AG

The driver assistance functions available on trucks produced by Daimler Truck AG are implemented in an Electronic Control Unit (ECU) called the Video Radar Decision Unit (VRDU). The VRDU receives data from sensors integrated into the vehicle and interacts with other ECUs, which control the brakes and the power steering for example.

The VRDU contains an component which records information when its driver assistance functions are triggered. This component is called the Situation Analysis Recorder (SAR). The SAR reacts to predefined triggers, such as an imminent crash, and responds by recording a number of parameters about the situation. These parameters are recorded at multiple points in time following the activation of the trigger. By regularly recording data into a ring buffer with fixed size, it is even possible to reconstruct a small amount of historic data. Event records themselves are also stored in a ring buffer, so that new event records overwrite records of older events.

2.2.2 Legal Considerations

This section outlines regulation which is relevant to event data recording in commercial vehicles. Since regulations vary widely between countries, only regulations applying to European markets, and German markets in particular, are considered.

Product Liability

ADAS can contribute greatly to safer road travel. Despite this, systematic failures in those systems could render them useless or even increase the risk of accidents. If such a systematic failure is discovered in a product, the manufacturer must recall the affected product and provide a fix. Additionally, the manufacturer could be liable for damages resulting from the failure.

To protect themselves from costly product liability claims, manufacturers include event recording systems into their vehicles [ADA19]. Manufacturers can use recorded event data to prove that an ADAS was operating normally before and during an accident, indicating that the accident did not result from a systematic failure in the ADAS.

Additionally, recorded event data may be used to observe product behavior in the field. In this case, a comparison of the observed product behavior to expected product behavior can serve as an early warning signal.

Data Protection

Data protection regulations are important to consider when implementing event data recording. The relevant regulation for European markets is EU regulation 2016/679, the General Data Protection Regulation (GDPR) [GDPR].

The GDPR places requirements on the processing of personal data. Here, personal data refers to "any information relating to an identified or identifiable natural person" [GDPR] and processing refers to any activity that touches such personal data. In the context of automotive event data, Vehicle Identification Numbers (VINs), which are commonly used to identify vehicles, are considered personal data. Therefore, VINs, and any data collected or stored with reference to a VIN, must be processed according to the GDPR.

Additionally, the GDPR defines the terms data subject and controller [GDPR]. The data subject is the person whose personal data is processed. The controller is the entity that defines the data processing.

The GDPR requires a legal basis for any processing of personal data [GDPR]. Possible legal bases include consent of the data subject and legal obligations of the controller.

Further, the GDPR requires all processing of personal data to be for an explicit purpose, which must be specified [GDPR]. Data collected for one purpose may not be used for another, incompatible purpose [GDPR].

The GDPR also grants data subjects extensive rights with regards to their data processing. For example, the controller must inform the data subject about any processing which will occur with the subject's personal data, the legal basis for that processing, and the purpose of that processing when collecting that data [GDPR]. Additionally, the controller must inform the data subject of any changes that occur during the processing of the subject's personal data [GDPR].

Controllers must also limit the personal data they process to the minimal amount necessary to achieve the specified purpose and delete any personal data as soon as the processing of the personal data is finished [GDPR].

Furthermore, controllers must ensure the integrity and confidentiality of personal data they process [GDPR]. For example, access to personal data must only be granted where necessary for fulfillment of the specified purpose.

Since the processing of the VRDU event records refers to the VINs of the affected vehicles, the event records must be treated as personal data. Therefore, the requirements of the GDPR must be fulfilled.

General Safety Regulation

As noted previously, the GSR requires EDRs in commercial vehicles from January 7, 2026. The GSR requires these EDRs to record a variety of parameters about the vehicle and its ADAS "shortly before, during and immediately after a collision" [GSR19]. The event records have to survive the collision [GSR19]. Additionally, the event records have to be protected against manipulation and must be anonymized [GSR19]. The GSR also forbids recording any information that could identify a specific vehicle, making special mention about the last four digits of the vehicle indicator section of the VIN [GSR19]. Further, all processing of event records must occur in accordance with the GDPR [GSR19].

It is important to note that the SAR component of the VRDU is not an EDR in the sense of the GSR. The first difference lies in the scope of the data recording. The GSR specifies that event records should contain data about all safety functions in a vehicle. In contrast, event records in the VRDU only contain data about a single ADAS or a small number of ADAS. The second difference is the definition of an event: While the GSR only considers actual collisions, the VRDU records data whenever its ADAS activate.

Nevertheless, requirements from data protection regulation, such as the GDPR, apply to the SAR component the same as to EDRs. Therefore, data protection requirements which the GSR places on EDRs can serve as guidance for further development of the SAR component.

Straßenverkehrsgesetz

The German Straßenverkehrsgesetz (StVG), which roughly translates to "road traffic regulation", requires a form of event data recording in vehicles equipped with high or full driving automation [StVG, §63a]. This recorder must record all handovers of control between the driver and the driving automation system. The event records must contain a timestamp and the vehicle coordinates as reported by a satellite navigation system.

The StVG is not yet applicable to the VRDU, which only performs partial driving automation. Therefore, event data recording as mandated by the StVG will not be considered in this thesis any further.

2.3 Data Pipelines

Data pipelines are chains of data-centric processes, where the output of a process can serve as input to other processes [MBO20]. Each process in a data pipeline performs some processing on the data which pass through that process. This processing can take different forms, such as data generation, data integration, or data storage.

Data generation is the first activity in a data pipeline [RBOW20]. A process, the data source, generates samples, which are fed into the data pipeline.

In reality, there are often multiple sources in any given data pipeline [RBOW20]. The data sets collected from those multiple sources need to integrated with each other. According to Golshan et al., data integration aims to provide a uniform query interface to multiple input data sets [GHMT17]. To provide this uniform query interface, the schemas of the constituent data sets must be harmonized into a global schema [GHMT17].

Data transformation also presents an important process in data pipelines [MBO20]. Examples of data transformations are converting unstructured data into semi-structured or structured formats or deriving new features from existing features [RBOW20].

Another component of data pipelines is data storage [RBOW20]. Data storage can occur at different stages and with different kinds of data [RBOW20].

One type of data storage are data lakes. They store large amounts of raw data in its original format, often an unstructured one [MT16].

Data warehouses present another type of data storage. In contrast to data lakes, data warehouses store structured, integrated, and aggregated data [HLV00; MT16].

Derived from the data stored in data warehouses, data marts are a third type of data storage [MK00]. Data marts enable data analytics by providing end users with tailored subsets of the data stored in a data warehouse [MK00].

Data analysis is another component of data pipelines [MBO20]. Often, this means applying machine learning models to the data set [MBO20].

Additionally, data pipelines may contain visualization components [MBO20]. These components aim to present data and data analysis results in an understandable manner [MBO20].

2.4 Data Analysis

As noted in Section 2.3, data analysis is an important part of a data pipeline. Numerous approaches to data analysis exist. The problem that data analysis is trying to solve defines which data analysis approaches are feasible [MRT18].

2.4.1 Supervised and Unsupervised Data Analysis

Data analysis problems are often distinguished between supervised problems and unsupervised problems [MRT18]. With supervised data analysis problems, a labelled data set is already available [MRT18]. This data set can be used to tune, select, and evaluate data analysis approaches. Classification problems are a common example of supervised data analysis problems [MRT18].

In unsupervised data analysis problems, no labelled data set is available [MRT18]. Therefore, unsupervised data analysis often is of a more exploratory nature. The lack of a labelled data set also complicates evaluating the performance of analysis approaches [MRT18].

2.4.2 Clustering

Clustering is a common class of problems in unsupervised data analysis [MRT18]. In clustering, the goal is to create groups of similar samples from a data set [ABKS99]. The similarity of samples can be judged using distance measures.

Ankerst et al. propose Ordering Points To Identify the Clustering Structure (OPTICS), a density-based algorithm, which can be used to cluster data sets [ABKS99]. OPTICS produces an ordering of the data set, which represents density-based clustering structure of the data set [ABKS99]. From this ordering, clusters can be extracted with variable local density parameters [ABKS99]. This improves on previous density-based methods, which

only performed clustering using a global density parameter [ABKS99]. Additionally, Ankerst et al. propose reachability plots, a visualization of the ordering produced by OPTICS.

2.4.3 Few-shot Learning

Supervised data analysis approaches often require large amounts of labelled training data [SS20]. Few-shot learning is an active research topic and aims to reduce the amount of required training data [WYKN20]. With few-shot learning, supervised data analysis approaches could be applied to problems where obtaining large amounts of training data is difficult or even impossible [WYKN20]. Smaller training data sets also speed up the training process [WZTE18].

Wang et al. propose dataset distillation as an approach to few-shot learning [WZTE18]. In their work, they were able to compress a data set containing 60,000 images into 10 synthetic training images, which achieve nearly the original performance when used as training data.

A special subclass of few-shot learning is less-than-one-shot learning, as proposed by Sucholutsky and Schonlau [SS20]. In less than one-shot learning, the learning task has to learn N classes from $M < N$ samples [SS20]. Sucholutsky and Schonlau were able to achieve this using soft labels.

2.4.4 Dynamic Time Warping

As noted in Section 2.4.2 on the facing page, clustering aims to group samples based on similarity, as judged by a distance measure. Dynamic Time Warping (DTW) can be used as a distance measure for time series [WK20]. DTW is a dynamic programming algorithm, which aims to eliminate nonlinear timing differences between two time series [SC78]. In general, DTW can be applied to any two sequences of possibly multi-dimensional points.

DTW constructs a mapping sequence between two sequences, which maps indices of the first sequence to indices of the second sequence [SC78]. This mapping sequence minimizes the sum of distances between points that are mapped to each other [WK20]. The mapping sequence also has to adhere to several conditions [SC78].

1. The first indices of both sequences are mapped to each other.
2. The last indices of both sequences are mapped to each other.
3. Indices are monotonically increasing.
4. Indices must not be skipped.

Additionally, an upper limit may be placed on the amount of warping between the two sequences [WK20]. The upper limit may either be expressed as an absolute number or as a percentage of the length of the time series [WK20]. Wu and Keogh point out

that an upper limit on the amount of warping improves accuracy [WK20]. Since an upper limit on the amount of warping restricts the search space, it may also speed up the computation [WK20].

3 Related Work

Several studies in the area of data pipeline management or in the context of automotive event data have been conducted. This chapter outlines a selection of those studies and their differences to this thesis.

Raj et al. conducted a study to identify which challenges practitioners face in data pipeline management [RBOW20]. Additionally, they developed a meta-model for data pipelines. They model processes as nodes and data flow as connectors between the processes. Nodes have capabilities, which represent activities they perform. Examples of node capabilities are data generation, data storage, and data labelling. Connectors also have capabilities. All connectors share the ability to transmit data, but they can have additional capabilities, such as authentication or validation. Finally, Raj et al. validated the developed meta-model against multiple data pipelines in a validation study. In contrast to this thesis, the challenges which Raj et al. identified were mostly of a technical nature.

Munappy et al. also studied the challenges in data pipeline management [MBO20]. They conducted interviews with 16 participants from 4 companies to identify those challenges. In these interviews, Munappy et al. identified infrastructure challenges, organizational challenges, and data quality challenges, all of which affect data pipeline management. They also identified opportunities stemming from the challenges. In contrast to their work, this thesis has a smaller scope, only studying one company. Further, this thesis contains a design component in developing an analysis approach for classification using small pre-labeled data sets.

A similar data analysis was the focus of a study conducted by Jurczyńska [Jur19]. In their work, they analyzed data from a system comparable to ABA. In contrast to the data set considered in this study, their data set already contained a comparatively large amount of labeled data. Furthermore, their data set consisted of multi-variate time series with fixed length and sampling time, which contain 132 signals. In contrast, the data set in this thesis contains significantly fewer signals and the time series are shorter and were sampled at a dynamic rate.

In their work, Johanson et al. presented a framework for data pipelines working with automotive telematics data [JBJ+14]. Their solution allows users to specify measurement tasks and assign these tasks to connected test vehicles. The test vehicles collect the required data and upload it back to the solution, either through a batch upload or through streaming. Users can then specify analytics tasks on the collected data. Johanson et al. developed an end-to-end data pipeline from data collection in vehicles to data analysis in Apache Spark from scratch, while this study focuses on extending an existing data pipeline. Additionally, they mainly worked with diagnostic readouts and passive bus monitoring data. This study, in contrast, works with event records collected from ADAS activations.

Jovanovic et al. proposed Quarry, a big data integration platform [JNR+20]. The platform aims to support non-technical end users in data exploration and data integration, which it tries to automate. Additionally, Jovanovic et al. validated their approach on a use case with World Health Organization. In contrast to this thesis, their work focused on data integration. Also, their use case was health related, while this study works with event data from ADAS.

Pevec et al. also proposed a big data platform for the automotive industry [PVG+19]. Their focus was on developing a robust, scalable and fault tolerant platform, which can cope with the amount of data produced by big data applications. They also showcased the developed platform in two real-world use cases. In contrast to this thesis, their approach focused on developing a platform and did not study challenges for automotive data integration and data analysis pipelines more deeply. Pevec et al. also did work with data from passive bus monitoring in one of their use cases and with transaction data in their second use case. This thesis, on the other hand, designs a data analysis approach for automotive event data.

To summarize, no comprehensive study of challenges and requirements to data integration and analysis in the context of automotive event data has so far been conducted. Previous studies either focused on data analysis, not reporting on challenges for data integration and data analysis, or did not target data pipelines for automotive event data in particular. This thesis aims to fill this gap by combining research about challenges for data integration and data analysis with designing a concrete data analysis procedure for automotive event data from ADAS.

4 Study Design

This chapter outlines the research question I was trying to answer with this study. In addition, it describes the research process I followed to answer these research questions.

4.1 Research Questions

This section describes the research questions. Six research questions were defined for this study. They were split into three groups.

The first group of research questions asks about challenges when working with a data integration and analysis pipeline in different contexts.

RQ1.1 What challenges do companies face when implementing or extending a data integration and analysis pipeline?

RQ1.2 Which additional challenges do arise when analyzing event data in the context of automotive events?

RQ1.3 How do challenges that Daimler Truck AG faces differ from challenges in the context of automotive events?

The second group contains only one research question, which focuses on data analysis for the ongoing improvement of driver assistance systems.

RQ2 Which data analysis questions are suitable for supporting the improvement of active driver assistance systems?

The third group of research questions builds on RQ2 and is trying to answer one of the data analysis questions derived in RQ2. Based on the results of RQ3.1, a concrete analysis was developed. RQ3.2 evaluates this analysis.

RQ3.1 Which concrete algorithms exist to answer the identified data analysis question?

RQ3.2 How does a concrete implementation of one of the algorithms perform on the available event records?

4.2 Research Methods

This section describes the research process I used to answer the research questions defined in Section 4.1. Figure 4.1 gives an overview of that research process. I started by conducting a literature review (Section 4.2.1) and interviews (Section 4.2.2) to elicit challenges when working with data analysis pipelines. Based on the results of these two steps, a prototype of a concrete analysis was developed and evaluated on a subset of the available event data (Section 4.2.3).

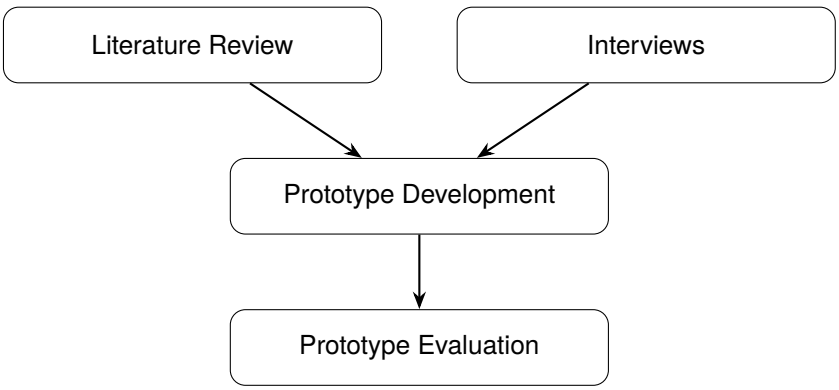


Figure 4.1: Research process

4.2.1 Literature Review

I conducted a rapid review [CPS20] to answer RQ1.1 and RQ1.2. This section will outline the process I followed during this review.

Process

I searched for papers using Google Scholar[1]. For RQ1.1, I used the query: `"data analysis" challenges "big data"`. For RQ1.2, I used the query: `"data analysis" challenges "big data" automotive`.

Only English papers were considered for the review. Originally, I also allowed German papers, however, this was not necessary since the most relevant results for each query were available in English. I also limited the review to papers that underwent peer review prior to publication. Further, I intended to exclude any paper that was considered for answering RQ1.2 from consideration for answering RQ1.1. However, no papers had to be excluded from RQ1.1 for this reason, since no overlap occurred.

[1] https://scholar.google.com

During the search, papers were sorted by relevance, as ranked by Google Scholar. Each paper was evaluated against the selection criteria outlined above. If a paper was excluded by the selection criteria, it was replaced with the next most relevant paper from the Google Scholar results.

Analysis

The resulting papers were analyzed through open coding in a constant comparative method [AHK11]. Each paper was analyzed line by line to identify challenges when working with a data pipeline. Identified challenges were highlighted using the "Comment" feature in Adobe Reader[2].

Whenever that analysis indicated a new challenges, this challenge was compared against a list of previously defined challenges to identify similarities. If sufficient similarity was found, the existing challenge was updated, otherwise the new challenge was added to the list. The list of challenges contained the code, a description of the challenge, and a list of papers supporting the challenge. A separate list of challenges was kept for each research question.

4.2.2 Interviews

To answer RQ1.3 and RQ2, I conducted semi-structured interviews [HA05] with stakeholders of the data analysis pipeline at Daimler Truck AG. This section describes the interview and analysis process.

Process

Potential participants were informed about the general topic of the study in an initial message. This initial message also informed the potential participants about the planned interview process. Finally, the initial asked potential participants whether they were willing to proceed with the interviews. When a potential participant agreed to the interview, a meeting was scheduled with them.

All interviews were conducted remotely through Skype[3] meetings. At the beginning of the meeting, the participant was informed about the purpose of the study and the interview process again. Afterwards, the interviewee had the chance to ask any questions. When the interviewee had no more questions, they were asked for their permission to record the interview. After the participant agreed to the recording, the recording was started and the interviewee was asked to repeat their permission so that it was captured in the

[2] https://acrobat.adobe.com/de/de/acrobat/pdf-reader.html
[3] https://skype.com

recording. Participant audio was captured using Audacity[4]. Audio for the interviewer was not captured for the first few interviews. Later the iOS voice memo app[5] was used to record interviewer audio.

During the interview, interviewees were asked some general questions about their background. They were also asked about their experiences with the data pipeline and prior analyses they performed on the data set considered in the study. A full interview guide is available in Appendix B on page 69. However, the questions presented in the guide are more of a general guidance and were adapted to the interviewee's previous answers. Some questions were skipped to better fit the background of the interviewee.

After the interview, the recording was transcribed using oTranscribe[6]. The transcript was then sent to the interviewee for their approval. This also gave the interviewee the chance to edit or remove statements from the transcript. After the interviewee's approval of the transcript the recording of the interview was deleted.

During the interviews with ADAS developers, possible analysis questions were also discussed. Based on this discussion, a data analysis question was defined.

Analysis

Interview transcripts were again analyzed through open coding using a constant comparative method [AHK11]. Each transcript was analyzed line by line to identify challenges for data pipeline management. Identified challenges were highlighted using the "Comment" feature in Microsoft Word[7].

The analysis of interviews followed the same process that was used for analyzing the literature review, which was described in Section 4.2.1.

In addition to open coding, axial coding was performed to identify connections between the discovered challenges [AHK11]. When a new challenge was discovered or a transcript indicated a relationship between two challenges, the use of challenges involved in the connection was checked against the interview transcripts. This ensured that the role of the challenge in the connection matched its use in the interview transcripts. When the use of the challenge in the interview transcripts did not match, the possible connection was discarded. Otherwise, the possible connection was checked against a list of previously defined relations to identify similarities. Only if no sufficient similarity was found, a new connection was added to the list.

[4] https://www.audacityteam.org
[5] https://support.apple.com/en-us/HT206775
[6] https://otranscribe.com/
[7] https://www.microsoft.com/de-DE/microsoft-365/word

4.2.3 Analysis Prototype

As noted previously, a part of this thesis is developing a concrete data analysis. This section outlines the approach taken for this development. It also outlines how the developed prototype was evaluated.

Data Sets

Two data sets were created for development and evaluation of the prototype. The first data set is the testing data set, which was randomly sampled from all available event records. The testing data set was used to evaluate the performance of the prototype.

A smaller validation data set was used for development and parameter tuning of the prototype. This data set was randomly sampled from the testing data set.

Additionally, the validation data set contained three event records where the situation surrounding the event is known. Two of those event records describe similar situations, one record describes a significantly different situation. These event records were used during development to tune the distance metric.

Another event record with known situation was included in the testing data set but not in the validation data set. This event record was used to check whether the developed prototype and distance metric generalize to new situations.

Development

Based on the data analysis question defined in the interviews, I researched possible data analysis approaches that could be used to answer the analysis question. For these approaches, feasibility and possible consequences were estimated. Based on these estimates, I decided to answer the analysis question through clustering of the event data.

The prototype was developed in an iterative manner with frequent feedback from ADAS developers. During development, silhouette coefficients were used to estimate the performance of the algorithm on the validation data set. Silhouette coefficients were also used to optimize the hyper-parameter selection for the clustering approach.

Evaluation

The performance of the developed prototype was then evaluated on the testing data set. As a baseline, the clustering was performed using Manhattan distance [Hor85] as a distance measure. This baseline was compared against clustering using DTW as a distance measure.

For both distance measures, silhouette coefficients [Rou87] were used to evaluate the performance of the clustering algorithm. The silhouette coefficient s for a single sample can be computed as

$$s = \frac{b - a}{\max(a, b)}$$

where a is the mean distance between the sample and all other samples in the same cluster and b is the mean distance between the sample and all other samples in the next-nearest cluster. Therefore, silhouette coefficients are only applicable, if at least two clusters are identified. The values for the silhouette coefficient range from -1 to 1, with higher values indicating a better cluster assignment. The mean silhouette coefficient over all samples was calculated to provide a quick overview over the clustering performance. Mean silhouette coefficient was chosen over other metrics, such as the Caliński-Harabasz [CJ74] index or the Davies-Bouldin [DB79] index, since silhouette score does not require the computation of cluster centroids. Additionally, ground truth labels were unknown, therefore, evaluation metrics that rely on ground truth labels, such as the adjusted rand score [HA85] or the V-Measure [RH07], were inapplicable.

Additionally, the number of clusters and the number of noisy samples were evaluated. The cluster assignment of the four events discussed previously was also analyzed.

Furthermore, for each cluster, the five event records with the highest silhouette score were analyzed individually in order to characterize the cluster. I performed an initial analysis of these event records. The results of this analysis were then validated with domain experts.

The agreement between both clustering variants was also analyzed using Adjusted Mutual Information (AMI) [VEB09] and a contingency matrix. AMI measures the agreement between a ground truth cluster assignment and a cluster assignment produced by a clustering algorithm. The agreement score ignores permutations in cluster assignments. It produces values from 0 to 1, with higher values indicating higher agreement. As noted previously, a ground truth cluster assignment is not known. However, AMI is symmetric and can therefore also be used to measure the agreement between cluster assignments produced by two different clustering algorithms, like in this use case. AMI is also normalized against chance, producing values close to 0 for randomized cluster assignments.

A contingency matrix can also be used to visualize the agreement between two clustering algorithms. In this use case, rows corresponded to clusters under the baseline distance measure and columns corresponded to clusters under the DTW distance measure. As such, the element in row i and column j gave the number of event records that were assigned to cluster i under the baseline distance measures and to cluster j under the DTW distance measure. In case of significant disagreements in the contingency matrix, a sample of these event records was also analyzed in an attempt to identify the reason for the disagreement. This analysis was also performed by myself and the results were validated with domain experts.

5 Results

In this chapter, I will report the results from the literature review (Section 5.1) and from the interviews (Section 5.2). Additionally, the architecture and implementation of the prototype of a concrete data analysis will be described (Section 5.3).

5.1 Literature Review

As noted previously, I conducted a rapid review to identify common challenges for data analysis endeavors. The following sections discuss the results from this rapid review. Results are split into two parts, the first part describing challenges for data pipeline management in general (RQ1.1), the second part describing additional challenges for data pipeline management in automotive contexts (RQ1.2). Both research questions were defined in Section 4.1 on page 27.

5.1.1 General Challenges for Data Pipeline Management

In total, the literature search identified five studies, four of which were judged as relevant. These four studies were included in the literature review to collect challenges for data pipeline management in general. The collected challenges were used to answer RQ1.1. A list of these studies is available in Appendix A.1 on page 67.

Table 5.1 gives an overview over the identified challenges and the studies that support them. In the following, these challenges will be described in more detail.

All relevant studies mentioned *heterogeneity*, also called variety, as a significant challenge for big data analytics [FHL14; IS15; LJ12; TEK+16]. In many cases, big data sets are created by combining data from multiple sources [FHL14]. Typically, these data sources do not conform to a single schema, introducing small differences between raw data collected from different data sources. These small differences introduce heterogeneity when data sets from different data sources are combined. Heterogeneity can also result from evolution of a single data source over time [FHL14]. Heterogeneous data sets present a challenge for traditional data storage systems [IS15] and analysis approaches [FHL14]. Heterogeneity is one of the inherent properties of big data [IS15].

The studies also agreed that *data volume*, another inherent property of big data, is a challenge for big data analysis [FHL14; IS15; LJ12; TEK+16]. Data volume refers to the number of individual samples in a data set. Data sets with more samples require more efficient algorithms [FHL14; TEK+16]. Even with optimal algorithms, large data volumes may

Challenge	Supporting studies
Heterogeneity	Fan et al. [FHL14], Iqbal and Soomro [IS15], Labrinidis and Jagadish [LJ12], and Tetko et al. [TEK+16]
Data volume	Fan et al. [FHL14], Iqbal and Soomro [IS15], Labrinidis and Jagadish [LJ12], and Tetko et al. [TEK+16]
Data quality	Fan et al. [FHL14], Labrinidis and Jagadish [LJ12], and Tetko et al. [TEK+16]
Understandability	Labrinidis and Jagadish [LJ12] and Tetko et al. [TEK+16]
Data protection	Labrinidis and Jagadish [LJ12]
Data sharing	Tetko et al. [TEK+16]

Table 5.1: Overview of general challenges for data pipeline management

make single core processing infeasible, creating a need for efficient parallelization approaches [FHL14; TEK+16]. Large scale parallel processing poses new challenges for data analysis and computing infrastructure [FHL14].

Three studies mentioned low *data quality* as a challenge for data pipelines [FHL14; LJ12; TEK+16]. For example, any step of data processing in a data pipeline may introduce noise into the data [FHL14; TEK+16]. Further problems with data quality are low information density, erroneous data, and incomplete data [LJ12]. Handling low data quality requires two things. First, low quality data must be identified. This needs approaches to identify specific issues with data quality. Second, after the issues leading to low quality data have been identified, remediation techniques are necessary to fix the identified issues. For example, redundant samples can be filtered to increase information density [LJ12]. Any attempt to improve data quality must be evaluated to prove its impact [LJ12]. Otherwise, the attempt might actually decrease data quality, for example by discarding interesting information [LJ12].

Two studies also reported *understandability* as a challenge of data analysis [LJ12; TEK+16]. In data driven organizations, decision makers use data analysis results to inform their decisions [LJ12]. For this purpose, they must be able to understand the results presented to them and the analysis which produced these results [LJ12]. Visualization techniques present one approach which increases understandability [TEK+16].

Further, Labrinidis and Jagadish mentioned *data protection* as a challenge [LJ12]. As mentioned previously, as soon as a data set contains personal information, data protection regulations apply and special care must be taken when handling the data.

Closely related to data protection, is *data sharing*, which Tetko et al. identified as a challenge [TEK+16]. Big data analytics in organizations is limited by the amount of data these organizations can collect [TEK+16]. To improve big data analyses, organizations grow their data sets by sharing data with other organizations [TEK+16]. How to share and

integrate these data sets efficiently and securely is a technical challenge [TEK+16]. If personal information is involved, data sharing can be further complicated by data protection regulation, such as the GDPR.

5.1.2 Challenges for Data Pipeline Management in Automotive Contexts

In total, the literature search identified five studies, four of which were judged as relevant. These four studies were included in the literature review to identify additional challenges for data pipeline management in automotive contexts, answering RQ1.2. A list of these studies is available in Appendix A.2 on page 67.

Since RQ1.2 asks about additional challenges in the context of automotive event data, I will not repeat challenges that were already reported in Section 5.1.1. Table 5.2 gives an overview over the identified challenges and the studies that support them. In the following, these challenges will be described in more detail.

Challenge	Supporting studies
Real-time requirements	Ge and Jackson [GJ14], Johanson et al. [JBJ+14], and Syafrudin et al. [SAFR18]
Mobility	Johanson et al. [JBJ+14] and Pevec et al. [PVG+19]
Bandwidth	Johanson et al. [JBJ+14] and Pevec et al. [PVG+19]
Communication interfaces	Johanson et al. [JBJ+14] and Pevec et al. [PVG+19]
Security	Ge and Jackson [GJ14] and Johanson et al. [JBJ+14]
Scalability	Johanson et al. [JBJ+14] and Pevec et al. [PVG+19]
Functional safety	Johanson et al. [JBJ+14]
Data format	Johanson et al. [JBJ+14]
Time-series data	Johanson et al. [JBJ+14]

Table 5.2: Overview of challenges for data pipeline management in automotive contexts

The main challenge, which was reported by three studies, are *real-time requirements* [GJ14; JBJ+14; SAFR18]. In many big data applications in automotive contexts, data must be processed in a limited time frame after it is produced [JBJ+14]. These real-time requirements can apply to processing performed inside the vehicle, but also to processing occurring outside the vehicle, for example in a cloud environment. Consequences for violating real-time requirements can range from the data losing its value to possible safety hazards [JBJ+14]. Therefore, automotive big data applications must fulfill their own real-time requirements and also ensure that they do not cause other processes to miss their real-time requirements.

Another challenge in automotive big data applications is the *mobility* of vehicles [JBJ+14; PVG+19]. This mobility complicates data collection. Different approaches exist to still enable data collection, such as temporary storage or wireless communication [JBJ+14]. However, these approaches incur additional challenges themselves.

When performing data collection via wireless communication, *bandwidth* presents a challenge [JBJ+14; PVG+19]. In many cases, high bandwidth wireless communication is costly or not available at all [JBJ+14]. Additionally, any available bandwidth must be shared between multiple applications in a connected vehicle. Bandwidth restrictions can be addressed by different means, such as data compression or aggregation [JBJ+14]. For automotive big data applications without real-time requirements, reducing the communication frequency might also be a viable solution.

Communication interfaces present another challenge for automotive big data applications [JBJ+14; PVG+19]. Any data flow between two components in a data pipeline requires that both components have compatible communication interfaces [PVG+19]. Incompatibilities can occur on different levels, such as physical connectors, communication protocols, or data formats.

Another challenge for automotive big data applications is *security* [GJ14; JBJ+14]. Automotive big data applications must prevent unauthorized third parties from gaining access to the vehicle [JBJ+14]. The data processed by these applications must also be protected, for example to protect personal information [JBJ+14] or the competitive advantage of the organization performing the processing [GJ14].

The large volume of big data makes *scalability* another challenge for automotive big data applications [JBJ+14; PVG+19]. The infrastructure supporting a big data pipeline must support scaling its storage and processing capabilities to meet the demands of a big data application [JBJ+14]. However, scaling also introduces new challenges into the data pipeline, for example in fault tolerance.

Closely related to real-time requirements, *functional safety* also presents a challenge for automotive big data applications [JBJ+14]. Automotive big data applications must ensure that they do not endanger vehicle occupants, either directly or indirectly.

Another challenge, related to the need for compatible communication interfaces, are *data formats* [JBJ+14]. Automotive big data applications use a variety of data formats for communication and storage [JBJ+14]. Many of these data formats are proprietary or custom built. Therefore, these data formats are not natively supported by off-the-shelf big data analysis tools, so custom parsers must be implemented before data can be used for analysis [JBJ+14].

Automotive big data applications often produce *time series data*, which presents another challenge for big data analysis, since it requires specialized analysis techniques [JBJ+14].

5.2 Interviews

After the literature review, I conducted interviews with stakeholders at Daimler Trucks AG to elicit challenges they face in their work with the data pipeline. As such, the interviews aim to answer RQ1.3.

In total, six stakeholders agreed to the interview. With the exception of two interviews, all interviews stayed within the expected time frame of 30 minutes. The two interviews that ran long were with ADAS developers, who were asked additional questions to answer RQ2.

Table 5.3 lists all interviewees, their role, and their experience in the role. As can be seen from the table, there is a significant difference in the experience between ADAS roles and big data roles. This highlights that many non-IT companies are still in the process of leveraging the potential of their data.

Participant	Role	Experience (years)
P1	Data onboarding	2
P2	ADAS development	15
P3	ADAS development	10
P4	Data governance	2
P5	Data engineering & data science	1
P6	Data engineering & data science	3

Table 5.3: Overview of interview participants

5.2.1 Challenges

The interviewees identified multiple challenges for their work with the data analysis pipeline in the interviews. These challenges can be broadly grouped into three categories: technical challenges, organizational challenges, and legal challenges. The categories are not mutually exclusive, some challenges fit two or even all three categories.

Table 5.4 gives an overview over the identified challenges and the participants who reported them. The rest of this section will describe these challenges in more detail.

All interviewees agreed that *capacity constraints* present a challenge for their work. These constraints can apply to any resource whose supply is limited. In a technical context, resources could be processing power, memory, or disk storage. From an organizational perspective, resources could be developer time or money. In most cases, there are trade-offs between the different resources. For example, money can be used to buy additional processing power in a cloud environment.

Challenge	Reporting participants
Capacity constraints	P1, P2, P3, P4, P5, and P6
Limited Storage	P2 and P3
Data format	P1, P2, P5, and P6
Dimensionality	P2, P3, P5, and P6
Volume	P2 and P3
Communication interfaces	P1, P3
Expert knowledge	P2, P3, P4, P5, and P6
Availability	P1, P2, P3, P5, and P6
Perspectives	P4
Role of ADAS developers	P2, P3, P5 and P6
Data protection	P1, P2, P3, P4, and P6

Table 5.4: Overview of challenges for data pipeline management identified in the interviews

All interviewees also agreed that their limited developer time forces them to prioritize their efforts. However, this can delay tasks with a lower priority. Additional issues can arise when stakeholders prioritize tasks differently.

According to P2 and P3, limited storage, another form of capacity constraints, is a special challenge for the SAR. Since storage space is limited, only a small number of events can be stored. Additionally, only a small number of signals is recorded per event. Without frequent readouts, event records may be lost before they make it into the data analysis pipeline.

Data formats are another challenge, which was mentioned by four interviewees (P1, P2, P5, P6). Event records are not stored in a common data exchange or serialization format but in a proprietary format. This causes additional effort to understand the data format and implement a conversion utility. Since multiple programming languages are used to implement data analyses in the data pipeline, a separate conversion utility needs to be implemented for each language.

Four interviewees (P2, P3, P5, P6) mentioned high *dimensionality* as a challenge. In AI terminology, the dimensionality of a data point describes the number of features in this data point. In the context of the SAR, the dimensionality of an event record is the number of data items recorded for this event. The number of data items is the product of the number of recorded points times the number of signals recorded for each point. With each additional data item, understanding the data becomes more difficult. This slows down other processes, such as the implementation of format conversion utilities. High dimensionality is also a problem for many AI algorithms.

Two interviewees (P2, P3) also reported that the dimensionality of the recorded events is too low in some cases. Since the situations in which events occur can be almost arbitrarily complex, numerous signals are necessary to describe any single situation. However, due to limited storage capacity on the VRDU, not all of these signals can be stored in an event record. Therefore, a small subset of the available signals must be selected for recording. This signal subset may not be able to distinguish as many situations as the full set of signals. In this sense, low dimensionality limits the expressiveness of data.

Two participants (P2, P3) also identified the *volume* of events, that is the number of event records, as a challenge. This number will increase drastically with the advent of new connectivity solutions like the CTP. The volume of events complicates identification of relevant events in a data analysis.

Furthermore, two participants (P1, P3) identified *communication interfaces* as a challenge in data pipeline management. As noted previously, any two communicating processes in the data pipeline need a compatible communication interface to do so. If such an interface is not available, either one of the processes must be extended to include a compatible interface or an intermediary must be placed between the processes to translate the data.

Another challenge, which was reported by five interviewees (P2, P3, P4, P5, P6), is the *expert knowledge* required to work with the data pipeline. Working with data in a data pipeline requires expertise from different fields. For example, this knowledge could be data analytics methods, legal expertise, or domain knowledge about the data. No stakeholder can be an expert in all of these fields. Usually, they are an expert in one of the fields and have at least some knowledge in other fields. Still, they do not have all of the required knowledge. Therefore, efficient processes for collaboration and knowledge transfer are needed.

However, *availability* of the participants presents a challenge for knowledge transfer. Limited time and different priorities make it harder to transfer knowledge. For example, learning about the data format might be a high priority for a data scientist, while a domain expert who understands the data format already might see the task as a lower priority. In this situation, the task will be delayed or can produce results of a lower quality. Five participants (P1, P2, P3, P5, P6) reported that they already encountered similar situations due to limited availability.

P4 also named different *perspectives* of the involved parties as another challenge for efficient knowledge transfer and collaboration. Different subjects have different terminology and processes. These differences complicate knowledge transfer and collaboration. Third parties who are familiar with both subjects, such as P4, can help serve as translators and facilitate collaboration. According to P4, assigning the same expert to multiple similar collaboration efforts also improves the transfer of knowledge. In this way, experts can build familiarity with the other field, which they can transfer to subsequent collaboration efforts.

P4 also remarked that data scientists and domain experts might not see a reason to learn the basics of data protection regulation. This places additional burden on data governance experts and legal teams to ensure compliance with regulations. In contrast to P4's statement, all participants of the interviews were willing to learn about data protection regulation and ensure compliance.

The interviews also indicated different opinions about the role of ADAS developers in data analysis. P2 and P3, ADAS developers, regarded themselves as consumers of data and data analysis. On the other hand, P5 and P6, data scientists, aimed for a collaborative approach, where the ADAS developers are actively involved in data analysis and the evolution of the data pipeline. This difference could also present a challenge during data analysis efforts.

Five participants (P1, P2, P3, P4, P6) agreed that *data protection* requirements present a challenge for their work in the data pipeline. When the SAR event records are collected, they are attached to a VIN, therefore the records must be considered personal information. This means any data processing of these event records must honor the GDPR.

P4 explicitly mentioned another challenge in the context of data protection. As noted previously, the GDPR expects the controller to inform the data subject of any processing occurring with their data. In the context of commercial vehicles, the manufacturer, in this case Daimler Truck AG, seldom has direct access to the data subject, in this case the driver. This makes it harder for manufacturers to comply with the requirements of the GDPR.

On the other hand, P6 regarded strict data protection as a competitive advantage. They said respecting and protecting users' privacy will build trust. In the long term, this trust will probably increase business and users might be more willing to share data with organizations they trust.

5.2.2 Data Analysis Questions

With P2 and P3, I also discussed possible data analysis questions during the interviews. This identified the following two possible data analysis questions.

The first question is identifying similarities in the situations which surround events. Following the definition of Ulbrich et al., "a situation is the entirety of circumstances, which are to be considered for the selection of an appropriate behavior pattern" [UMR+15]. This means a situation describes the system and its environment, including the behavior of any other vehicles involved in the event, at the time when the event is recorded. Identifying similarities in the situations can help focus developer attention on frequently occurring situations. For example, frequent situations with a high rate of unnecessary emergency brakes can be prioritized. Information about the frequency of certain situations can also help with prioritization during testing.

The second possible data analysis question is, why emergency brakes are aborted. As explained in Section 2.1, the driver can override decisions made by ADAS in vehicles designed by Daimler Truck AG. This data analysis question aims at identifying patterns in the data that indicate whether a driver will override the emergency brake. This can also help development of ADAS by highlighting indicators for false positives.

5.3 Analysis Prototype

This section describes the design and implementation of the concrete analysis that was developed as part of this thesis.

5.3.1 Data Analysis Question

As noted previously in Section 4.2.2, I also discussed possible data analysis questions with P2 and P3 during their respective interviews. We agreed on identifying situations in which ABA events occur. For this purpose, similar situations should be grouped in some manner. As outlined above, this can help focus developer attention.

The ABA event records were chosen as analysis data set for two reasons. First, ABA is a mature system that has been in the market since 2006. Therefore, a large number of event records has been collected already.

Second, ABA records are already integrated into the data analysis pipeline. Therefore, there is no need to develop a parser for a custom format.

5.3.2 Data Analysis Approach

Since the data analysis question asked for grouping event records based on similarity, two analysis approaches came to mind. The first is classification, a supervised approach. The second approach is clustering, an unsupervised approach.

Because almost no labeled data was available for the ABA event records, unsupervised data analysis through clustering was chosen. However, where label information is available, it can be used to characterize the cluster if the clustering approach proves successful.

To answer RQ3.1, I compared different clustering algorithms and analyzed their suitability for the data analysis task. Table 5.5 gives a short overview over the considered algorithms. In the following, I will describe the algorithms in more detail. I will also decide on a single clustering algorithm to use in the concrete analysis and will justify this decision.

Algorithm	Parameters
K-means	number of clusters
DBSCAN	maximum neighborhood distance, minimal number of samples per cluster
OPTICS	minimal number of samples per cluster

Table 5.5: Comparison of clustering algorithms [PVG+11]

K-means clustering is a well known clustering algorithm, which splits samples into a configurable number of clusters. Each cluster is described by a centroid. K-means expects clusters to be convex and isotropic. This means clusters should be roughly spherical for optimal performance.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a density-based clustering algorithm, which defines clusters as areas of high density separated by areas of lower density. In DBSCAN, the maximum distance between two points in the same cluster as well as the minimal number of samples per cluster can be configured. DBSCAN does not place any assumptions on the shapes of clusters.

OPTICS, which has already been described in Section 2.4.2, is another density-based clustering algorithm. In contrast to DBSCAN, only the minimal number of samples per cluster needs to be configured. The maximum distance between two points in the same cluster is automatically adjusted based on the density.

I decided on using OPTICS as the clustering algorithm, because it does not place any assumptions on the shape of the clusters. Additionally, the number of clusters does not need to be configured. I decided against K-means clustering, since the scikit-learn implementation does not support custom distance metrics. I chose OPTICS over DBSCAN, since configuring the maximum distance between samples in a cluster is not necessary for OPTICS. Additionally, the OPTICS implementation of scikit-learn provides a method to extract a DBSCAN like clustering from the results of OPTICS clustering[1].

Another part of the clustering approach is how to measure the similarity of the event records. Usually, the similarity is measured using a distance measure. However, the ABA event records can be regarded as a multivariate time series with a dynamic sampling rate. Traditional distance measures, such as the Manhattan distance [Hor85], do not perform well on this kind of data. Therefore, I decided to use DTW as a distance measure for the clustering algorithm.

5.3.3 Architecture

As noted previously, there already is a large number of events available. However, there also was a number of manual readouts, which contained event records that were not yet integrated into the data pipeline.

Originally, the event records were collected through XENTRY. Figure 5.1 shows how these event records were processed into the data pipeline. The data pipeline collects readouts from the XENTRY system. Among other data, these readouts contain ABA function data items. These items are processed by a parser component to produce readable ABA event records.

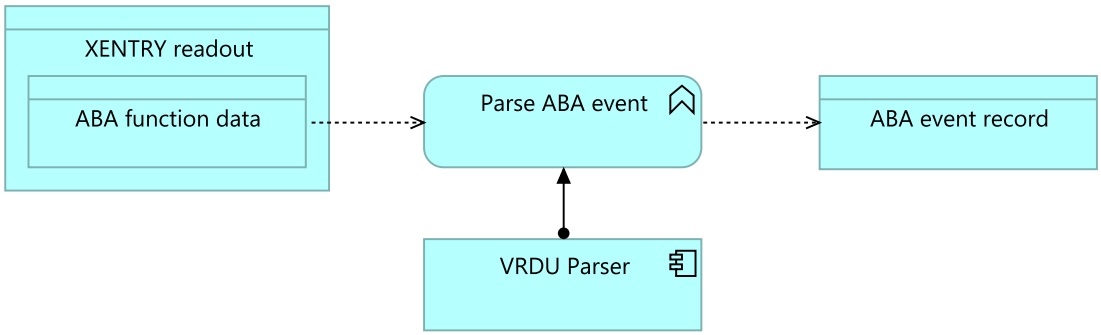


Figure 5.1: Data flow of ABA events from XENTRY readouts into the data pipeline (ArchiMate notation [TOG19])

[1] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.cluster_optics_dbscan.html

There is also a small set of manual readouts, which also contain ABA event records. For most of these manual readouts, the situation surrounding each ABA event is already known. As such, the event records from the manual readouts could help identify the situation represented by a cluster. Therefore, the data pipeline was extended to allow reading of ABA event records from manual readouts. The architecture is similar to that for XENTRY readouts. Manual readouts are uploaded into a staging directory. From there, the parser component processes the contained ABA function data items into ABA event records. The ABA event records produced from different readout sources are stored separately. A later step integrates the event records from different sources and performs deduplication.

5.3.4 Implementation

The analysis was implemented in Azure Databricks[2], which is based on Apache Spark[3]. Data processing was performed in pandas[4]. The tslearn library [TFV+20] was used for computation of DTW and preprocessing. The clustering was implemented in Python using the scikit-learn library [PVG+11].

The clustering is performed in the following four phases. Each phase will be explained in more detail later.

1. Event records are loaded and transformed into time series
2. If configured, the data is preprocessed.
3. A distance matrix is computed.
4. Clustering is performed on the distance matrix.

Data Loading

In the first phase, data is loaded from the data warehouse. As noted in Section 2.2.1, each event record contains a number of signals collected at multiple points during the event. However, not all points are available for all events. For example, if the driver aborts a haptic brake, no emergency brake will be triggered, therefore, no data will be collected at the point of the emergency brake. Additionally, not all signals are collected at all points. The first phase handles these situations by discarding points where no data has been collected and adding placeholders for signal values to points where those signals were not collected. In addition, the first phase transforms the data set from a tabular format into a time series format.

[2] https://azure.microsoft.com/de-de/services/databricks/
[3] https://spark.apache.org/
[4] https://pandas.pydata.org/

Preprocessing

The second phase performs three preprocessing steps, two of which are optional. The first step is resampling the time series. As mentioned before, the event records collected by the SAR have a dynamic sampling rate. During resampling, this sampling rate is converted to a fixed sampling rate and signal values are linearly interpolated. To evaluate the impact of different sampling rates on the clustering performance, resampling is optional.

The second preprocessing step is computation of derived signal values, such as relative acceleration or relative velocity. Additionally, at this point, a subset of signals on which to perform the clustering is selected and any points for which any signal values are still missing are discarded. In the clustering, relative signals were used to identify similar situations across different contexts. For example, situations where the leading vehicle is rapidly decelerating, such as during an emergency braking, should be clustered together, regardless of whether they occur on a highway or a country road. For the first approach, four signals were used in the clustering: the lateral deviation between the predicted vehicle course and the object d_{dev}, the time to collision ttc, the time to stop for the own vehicle tts_e, and the time to stop for the object tts_o.

The third step is rescaling. In rescaling, the time series are transformed so that each time series has mean 0 and standard deviation 1. With these transformations, the absolute value of the time series becomes less important, while trends can be more clearly identified. Rescaling is also optional, so that its impact on the clustering performance can be evaluated.

Distance Matrix Computation

In the third phase, a distance matrix is computed for the event records. This phase is necessary, as the scikit-learn clustering algorithms are not able to handle missing values in the feature matrix. Unfortunately, the size of the distance matrix limits the amount of events that can be processed to about 10,000.

As outlined in Section 4.2.3, clustering is performed using two different distance measures. Therefore, the distance measure used for constructing the distance matrix is selectable. Either the baseline distance measure, Manhattan distance, or the DTW distance measure are available. Computation of the Manhattan distance was adapted to time series by discarding any points beyond the shorter of the two time series. For DTW, the amount of warping as a percentage of time series length is configurable.

Clustering

In the fourth phase, OPTICS clustering is performed on the distance matrix. All results are saved for further analysis and a preliminary analysis is conducted. This preliminary analysis reports the number of identified clusters, the number of event records discarded as noise and the silhouette score of the clustering.

6 Evaluation

This chapter describes the evaluation of the prototype of a concrete analysis developed in the previous section. It also describes the data sets used for the evaluation and the parameter tuning approach that was conducted prior to the evaluation.

6.1 Data Sets

As mentioned in Section 4.2.3, two data sets were created for development and evaluation of the prototype. The testing data set of 1,000 event records is a subset of the integrated data set described in Section 5.3.3. The validation data set of 100 event records is a subset of the testing data set. As noted previously, the validation data set contained three selected event records, the testing data set contained four selected events. The remaining event records for each data set were randomly sampled.

6.2 Parameter Tuning

Before the evaluation, the hyper-parameters of the prototype were optimized. For this purpose, the analysis was performed on the validation data set with different hyper-parameter combinations. The performance of the analysis for each hyper-parameter combination was judged using the mean silhouette coefficient. Since the DTW distance measure requires more hyper-parameters, all parameter tuning was performed for the DTW distance measure and the selected parameters were reused for the baseline distance measure.

As described in Section 5.3.4, the analysis has multiple configurable hyper-parameters, which are listed in Table 6.1 along with the values used for parameter tuning. `warping` is given as a percentage of the maximum time series length. `min_samples` is given as a percentage of the total number of samples.

Parameter	Description	Values
`sampling`	the sampling rate to use for resampling	None, 50 ms, 100 ms, ..., 1 s
`scaling`	whether the time series should be rescaled	True, False
`warping`	the maximal amount of warping for DTW	1 %, 5 %, 10 %, ..., 100 %
`min_samples`	the minimal number of samples per cluster	1 %, 5 %, 10 %, 15 %, 20 %

Table 6.1: Hyper-parameters and values used for parameter tuning

Parameter tuning was performed through an exhaustive grid search. Clustering was performed for all possible parameter combinations and the achieved mean silhouette coefficient recorded. Then, the parameter combination achieving optimal performance was used for the evaluation run. Table 6.2 lists the hyper-parameter values which were chosen for the evaluation run.

Parameter	Value
`sampling`	650 ms
`scaling`	False
`warping`	5 %
`min_samples`	5 %

Table 6.2: Hyper-parameter values used for evaluation

6.3 Results

After parameter tuning, the analysis was run on the testing data set. The analysis was run twice: first with the baseline distance measure and second with the DTW distance measure. This section describes the results of these evaluation runs.

In Section 4.2.3, three measures were defined, which give a quick overview over the performance of the clustering algorithm. Table 6.3 lists the values of these metrics for both the baseline and DTW distance measure. The silhouette score was calculated per event. The value in the table shows the mean over all events. For computing the silhouette score without noise, events classified as noise were discarded before calculating the mean over the events. The number of clusters includes the noise "cluster". In summary, the metrics show that clustering produced good results under either distance measure, with clustering performing slightly better under the DTW distance measure.

Metric	Baseline	DTW
silhouette score (all events)	0.65	0.94
silhouette score (without noise)	0.95	0.96
number of clusters	7	7
number of noisy samples	150	39

Table 6.3: Overview metrics for evaluation

To visualize the performance of the clustering algorithm, a reachability plot was created for each distance measure, which can be seen in Figures 6.1 and 6.2. In each plot, every cluster has been assigned a different color. Events classified as noise are shown in black. As can be seen from the plots, both approaches produce well separated clusters.

The reachability plots visualize the reachability distance for each event record in the data set. Intuitively, the reachability distance of an event record is the distance at which the event record becomes part of a cluster. In the plots, clusters form valleys, which are separated from each other by steep inclines. Additionally, there is a noticeable gap in the range from 10^4 to 10^8 in both plots.

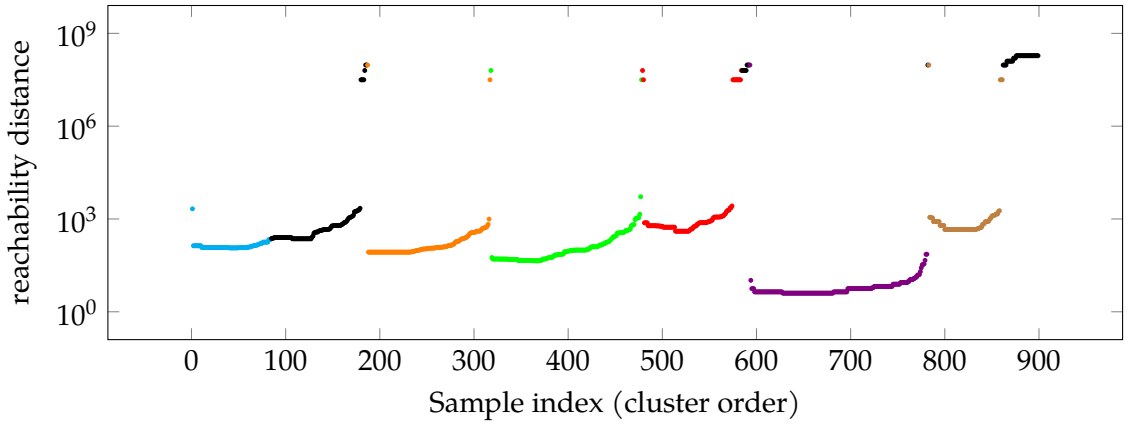


Figure 6.1: Reachability plot for the baseline distance measure

With knowledge of the contingency matrix, which will be discussed later, Figure 6.2 also shows that a number event records that were classified as noise under the baseline distance measure were assigned to cluster 0 under the DTW distance measure. One can also see that the order of clusters 2 and 3 has been swapped between the clusterings.

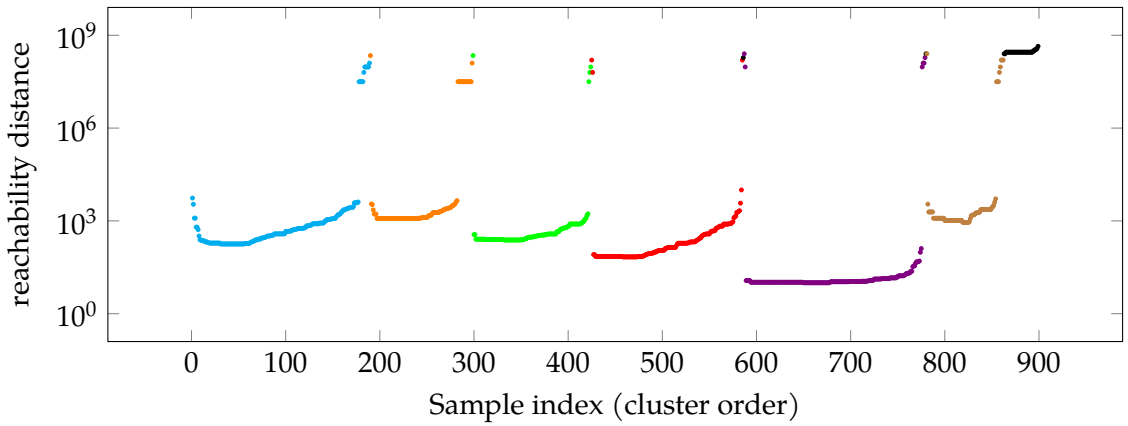


Figure 6.2: Reachability plot for the DTW distance measure

Additionally, Section 4.2.3 defined four events where the surrounding situation is known. Table 6.4 shows the expected and actual cluster assignment under both metrics. The table shows that the events were not clustered as expected. This indicates that the clustering approach did not perform as well as the overview metrics suggest.

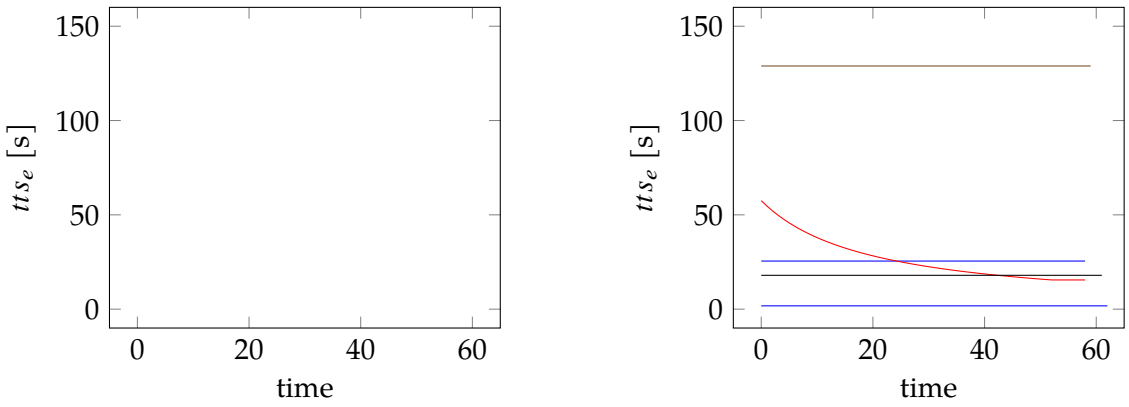
Figures 6.3 and 6.4 show traces for signals tts_e and tts_o respectively for the five samples with the highest silhouette coefficients from clusters 0 and 5. Considering all signals and clusters produces similar pictures. The plots indicate that the clustering is not based on

Event	expected	Baseline	DTW
A	x	3	1
B	x	2	3
C	not x	2	3
D	not x	4	4

Table 6.4: Cluster assignment for events in known situations

the signal values, but rather on the presence or absence of the signals. This is an artifact of how these signals were processed: Since scikit-learn cannot handle missing signal values directly, these missing signal values were replaced with a large placeholder. This placeholders, in turn, introduced large distances between event records if one of them is missing a signal. The distances stemming from missing signals are large enough to drown out smaller distances stemming from actual differences between the signal traces.

This can be clearly seen in Figure 6.3b. In addition to the wide distribution of absolute values for tts_e, there is one trace with a decreasing tts_e, corresponding to a braking ego vehicle. The other four traces show a constant tts_e. With a clustering based on similarity between situations, one would expect the first trace to be in a different cluster from the other four. That this is not the case, shows that the clustering is not based on the similarity of event records but on the presence of the signals.

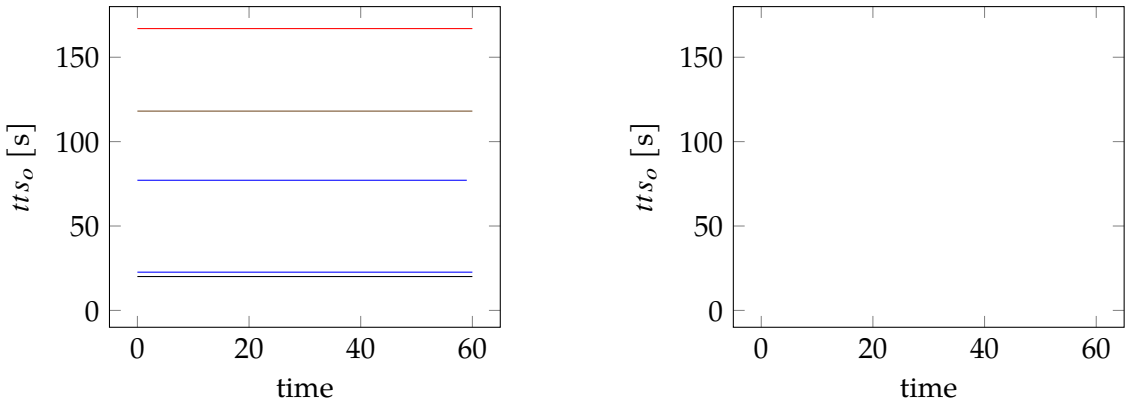


(**a**) Cluster 0

(**b**) Cluster 5

Figure 6.3: Comparison of traces for signal tts_e between clusters 0 and 5

Finally, I also analyzed the agreement between the clusterings produced by both metrics using the AMI score and a contingency matrix. The AMI score was 0.85, indicating a high agreement between both clusterings. This is confirmed by the contingency matrix, which shows clear matches between the clusterings produced by the baseline and DTW metrics. The only large disagreement between both clusterings affects samples classified as noise under the baseline distance measure. The majority of these samples has been assigned to the first cluster under the DTW distance measure.



(**a**) Cluster 0

(**b**) Cluster 5

Figure 6.4: Comparison of traces for signal tts_o between clusters 0 and 5

7 Discussion

This chapter discusses the results described in Chapters 5 and 6 on page 33 and on page 45. Section 7.1 discusses challenges for data integration and data analysis pipelines, while Section 7.2 discusses the evaluation of the analysis prototype. Finally, it also lists threats to validity and what measures have been taken to mitigate these threats (Section 7.3).

7.1 Challenges for Data Integration and Data Analysis

This section compares the issues identified in the literature review (see Section 5.1 on page 33) with those identified in the interviews (see Section 5.2 on page 37). It also describes approaches that could help to alleviate the challenges identified in the interviews.

7.1.1 Comparison

Table 7.1 lists the challenges for data pipeline management that were identified in the literature review or the interviews. It also shows whether a challenge was identified in the literature review, the interviews, or both. In the following, select challenges will be explained in more detail.

Several challenges were reported in the literature review as well as in the interviews. These are volume, data protection, communication interfaces, and data formats. Additionally, bandwidth restrictions, identified as a challenge in the literature review, can be regarded as a form of capacity constraints. Of these challenges, it is not surprising to see volume identified as a challenge in both the literature review and the interviews, since it is one of the defining characteristics of big data.

That data protection was identified as a challenge in the literature review and the interviews shows that it is a topic of interest for both research and industry applications. With organizations increasingly transforming into data-driven organizations, collection of personal data is also increasing. Therefore, these organizations have to comply with data protection regulation. When more organizations are collecting personal data, strong data protection also builds user trust and becomes a competitive advantage, as noted in the interviews.

The literature review identified communication interfaces and data formats as challenges for automotive big data applications. The interviews also identified these challenges. This indicates that those challenges are limited to the automotive industry in scope. Many big data analysis tools are cloud-based and as such use relatively young formats like JavaScript Object Notation (JSON), while many automotive systems have been in use

Challenge	Literature Review	Interview
Data volume	yes	yes
Data protection	yes	yes
Communication interfaces	yes	yes
Capacity constraints	(yes)	yes
Bandwidth	yes	(yes)
Data format	yes	yes
Heterogeneity	yes	no
Mobility	yes	no
Real-time requirements	yes	no
Functional safety	yes	no
Data sharing	yes	no
Data quality	yes	no
Understandability	yes	no
Security	yes	no
Scalability	yes	no
Time-series data	yes	no
Limited Storage	no	yes
Dimensionality	no	yes
Expert knowledge	no	yes
Availability	no	yes
Perspectives	no	yes
Role of ADAS developers	no	yes

Table 7.1: Comparison of challenges identified in literature review and interviews

for quite some time and rely on data formats available at the time of their inception. Additionally, embedded applications, which automotive applications are a special case of, have unique requirements, such as limited storage on ECUs, which reflect in the used data formats. This caused the development of specialized communication interfaces and data formats, which are difficult to integrate into modern data analysis tools.

Interestingly, heterogeneity was not reported as a challenge in the interviews. This could indicate that big data analysis tasks at Daimler Truck AG use less diverse data sources, thereby sidestepping heterogeneity as a challenge. However, it could also indicate that data engineers at Daimler Truck AG are sufficiently experienced in data integration, so that they no longer consider data integration a challenge.

Mobility and real-time requirements were also not reported as challenges in the interviews. A possible explanation is that the current use cases for big data analysis at Daimler Truck AG do not possess these requirements. Therefore, developers have not encountered the associated challenges yet.

Functional safety was also not reported as a challenge for automotive big data applications in the interviews. In the current big data analytics use cases at Daimler Truck AG, only data collection is performed on the vehicle. All other processing is performed in a cloud environment. Additionally, big data analysis cannot cause ADAS interventions into the driving of vehicles. Therefore, functional safety concerns do not apply to the data processing and functional safety is not perceived as a challenge for big data analytics.

The major challenge that was only identified in the interviews is the expert knowledge centralized in different departments. This challenge is probably not limited to a single company, such as Daimler Truck AG, but rather a part of the transformation process towards a data-driven organization. Efficient collaboration processes must be established to tear down knowledge silos and equip all stakeholders with necessary knowledge for their work. When companies gain more experience with big data analytics, they will probably also establish processes and guidelines that ensure this efficient collaboration.

7.1.2 Solution Approaches

Different approaches could help to alleviate the challenges identified in the interviews. This section presents some of those approaches.

One approach that could help decrease knowledge silos is a data catalog. A data catalog describes data sources available in a company. As such, a data catalog can aid data scientists in discovering data sources for data analysis. Additionally, a data catalog can contain detailed descriptions about the format of the data. That way, a data catalog could be used to understand data formats and data sources without requiring domain experts to be available for knowledge transfer. On the other hand, domain experts would be required to keep the information in the data catalog up to date, which places additional burden on them. However, with a large enough number of people attempting to learn about a data source, the time saved in knowledge transfers can offset the work required to maintain the data catalog. Currently, there is an effort to establish a data catalog at Daimler Truck AG.

The growing connectivity of vehicles can help automotive big data applications deal with limited storage available on ECUs. With connected vehicles, data can be collected from the vehicles more frequently, perhaps even in real time. This way, event records need to be kept on the ECU for a shorter time span, which means less event records must be kept at the same time. Therefore, more storage is available per event record, which can be used to record additional information about the event. This additional information, in turn, can improve data analysis and lead to new insights. At Daimler Truck AG, the Common Telematics Platform (CTP) provides a connectivity solution that could be used for continuous readout of connected vehicles. However, new connectivity solutions must comply with privacy regulations such as the GDPR.

Common data formats, like JSON, could be employed to ease the integration of new or evolving automotive data sources into the data pipeline. On the other hand, such data formats are often quite verbose, increasing the amount of memory required for data storage. With the limited amount of storage available on ECUs, this increase in data size could be unacceptable. However, other standardized formats, such as Measurement Data Format (MDF) [ASA19], could present an alternative.

Another alternative is the definition of the data formats in an Interface Description Language (IDL), such as Protocol Buffers[1], Cap'n Proto[2], or Apache Thrift[3]. AUTOSAR XML (ARXML) [AUT20] files, specified by the Automotive Open System Architecture (AUTOSAR) standard and widely used in the automotive domain, can also be used as an IDL. Using an IDL, the type and layout of data items can be described. Often, attaching additional metadata, like a range of valid values, to a data item is also possible. From a description in an IDL, code for serialization and deserialization in different languages can be generated. This would reduce the effort necessary to integrate a new data source or adapt the existing integration to changes in the data format, since the bulk of the parser code could be generated instead of needing to be implemented manually.

An IDL needs to fulfill some requirements to be applicable to the SAR use case. For example, it must support the data types used in the SAR, such as fixed point. Furthermore, the IDL should also be able to generate documentation for the data types.

The IDL description and generated documentation could also be integrated into a data catalog. From there, consumers of the data could retrieve the IDL description and use it to generate deserialization code for their use case. For this to work, the IDL must support code generation for the selected language.

7.2 Analysis Prototype

As noted in Section 6.3, the prototype of a concrete analysis did not produce the expected results. Instead of clustering event records based on their similarity to each other, the clustering was based on the presence of optional signals in the event record. This is not a suitable proxy for identifying the situation in which the event occurred.

In this section, I present alternative approaches which were investigated after the described approach failed to produce useful results.

7.2.1 Alternative Signal Combinations

First, I tried using different signal combinations to describe the event records for the clustering. Each signal combination was checked against the parameter combinations used for parameter tuning. The overview measures used in Section 6.3 were used to assess the

[1] https://developers.google.com/protocol-buffers
[2] https://capnproto.org/
[3] https://thrift.apache.org/

quality of the clustering results. Unfortunately, no signal combination resulted in a good clustering. Most signal combinations only identified a single cluster and classified 30 % to 60 % of the event records as noise. Consequently, the silhouette score was also low, with best signal combinations only reaching −0.08.

7.2.2 Feature Extraction

As a second approach, feature extraction was performed on the time series. In feature extraction, characteristics of the time series, such as extreme values, mean, or median, are computed. These characteristics, the features, were then used as the basis for clustering. In this approach, feature extraction was performed using the tsfresh library[4].

Since the clustering did no longer use time series data, the prototype was adapted accordingly. For example, the computation of a distance matrix became unnecessary and was removed. Also, clustering only used Manhattan distance as a distance measure, making the `warping` parameter obsolete. `rescaling` was also removed, since it was expected to interfere with extraction of the mean and variance features.

For the other parameters, the values listed in Table 6.1 were reused for parameter tuning. Parameter tuning was again performed using an exhaustive grid search. This determined that values of 100 ms for `sampling` and 1 % for `min_samples` performed best on the validation data set.

The tsfresh library provides three configurations for feature extraction, which differ in the features that are extracted. Feature extraction and clustering was performed and evaluated with all three configurations.

Additionally, feature selection was performed in an attempt to improve the clustering performance. In feature selection, the correlation of each feature with a target variable, here the assigned cluster, is evaluated and features with a low correlation are discarded. Three iterations of feature selection were performed, with each iteration performed on the clustering results of the previous iteration and the first iteration performed on the clustering results after feature extraction. Feature selection was again performed for all three feature extraction configurations.

However, clustering did not produce good results under any configuration for feature extraction. Feature selection did not improve this situation significantly. Table 7.2 shows the mean silhouette coefficient under all three feature extraction configurations. As can be seen from the table, best results were achieved by using minimal feature extraction configuration and without any feature selection. Nevertheless, the mean silhouette coefficient of −0.52 indicates significant overlap between the clusters. Additionally, 80 % to 90 % of event records were classified as noise.

Since feature extraction and feature selection did not produce promising results on the validation data set, I did not attempt a full evaluation of the approach on the testing data set.

[4] https://tsfresh.readthedocs.io/en/latest/index.html

	Feature extraction	Feature Selection		
Configuration		Iteration 1	Iteration 2	Iteration 3
comprehensive	-0.80	-0.71	-0.65	-0.64
efficient	-0.80	-0.70	-0.65	-0.62
minimal	-0.52	-0.52	-0.52	-0.52

Table 7.2: Mean silhouette coefficient for feature extraction and three iterations of feature selection

7.2.3 Synthetic Data Set

Finally, the clustering was attempted on a synthetic data set, which contained artificial events that were known to be similar. The events were generated through simulations runs with minor variations in starting parameters and noise application. With this data set I attempted to prove that DTW is able to identify similarities between events. If DTW is indeed able to capture the similarities, these similar event records would have to be clustered together.

Additionally, the synthetic data set contained a very small number of event records that were not similar to the other event records. These dissimilar event records were expected to be labelled as noise.

Clustering was performed using a variation of the prototype described in Section 5.3.4 on page 43. In contrast to the latter, this new prototype used different signals for the distance computation. These signals were the relative velocity v_{rel}, the relative acceleration a_{rel}, longitudinal distance between the own vehicle and the object d_x, and lateral deviation between the predicted vehicle course and the object d_{dev}. It was expected that these signals would be able to describe the situation surrounding an event.

In a first step, only the synthetic data set was used for the clustering. In this step, the prototype produced a single cluster containing the similar events and correctly identified the dissimilar events as noise.

In a second step, the synthetic data set was combined with the validation data set. With this combined data set, parameter tuning was performed following the approach described in Section 6.2 on page 45. The cluster assignment for the artificial event records in the clustering produced by the optimal hyper-parameter combination was also checked. Again, all similar events were assigned to the same cluster as expected. Table 7.3 shows the hyper-parameters that were chosen for the evaluation on the synthetic data set.

In a third step, an evaluation was performed for the clustering under baseline and DTW distance measures. For the data set, the synthetic data set was combined with the validation data set. Unfortunately, the clustering did not produce good results under either distance measure.

Parameter	Value
`sampling`	50 ms
`scaling`	False
`warping`	5 %
`min_samples`	5 %

Table 7.3: Hyper-parameter values used for evaluation on the synthetic data set

Under the baseline distance measure, two clusters were produced, while over 50 % of event records were discarded as noise. The mean silhouette coefficient was also rather low with 0.15. Figure 7.1 shows the reachability plot for the baseline distance measure on the synthetic data set. The plot highlights the large number of event records labeled as noise, denoted by black dots.

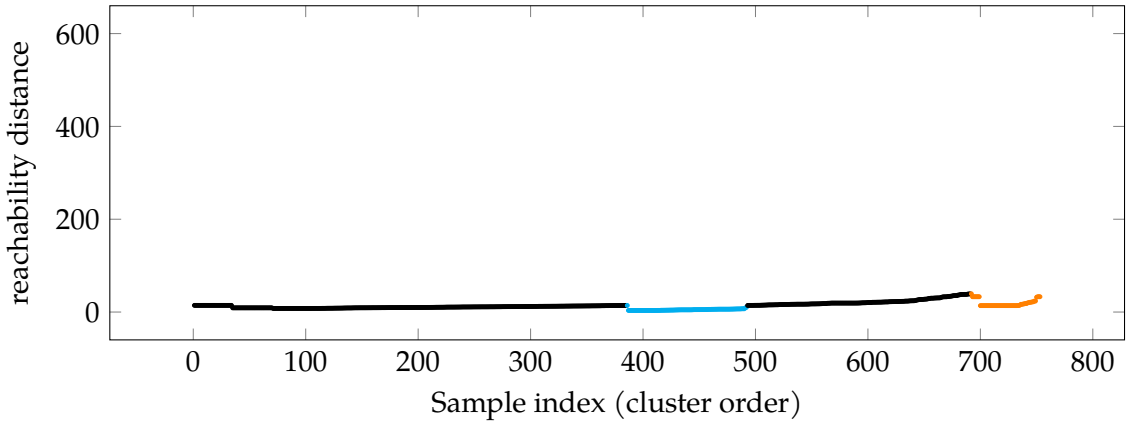


Figure 7.1: Reachability plot for the baseline distance measure on the synthetic data set

Under the DTW distance measure, only a single cluster was produced, but no samples were identified as noise. Perhaps, with the larger data set, some of the additional events are sufficiently similar to events from two different clusters that they cause these two clusters to merge. If that is the case, adding additional event records when moving to the testing data set could have caused the clustering algorithm to produce only a single cluster. Figure 7.2 shows the reachability plot for the DTW distance measure on the synthetic data set. In contrast to Figure 6.2, there is no empty horizontal band separating the clusters.

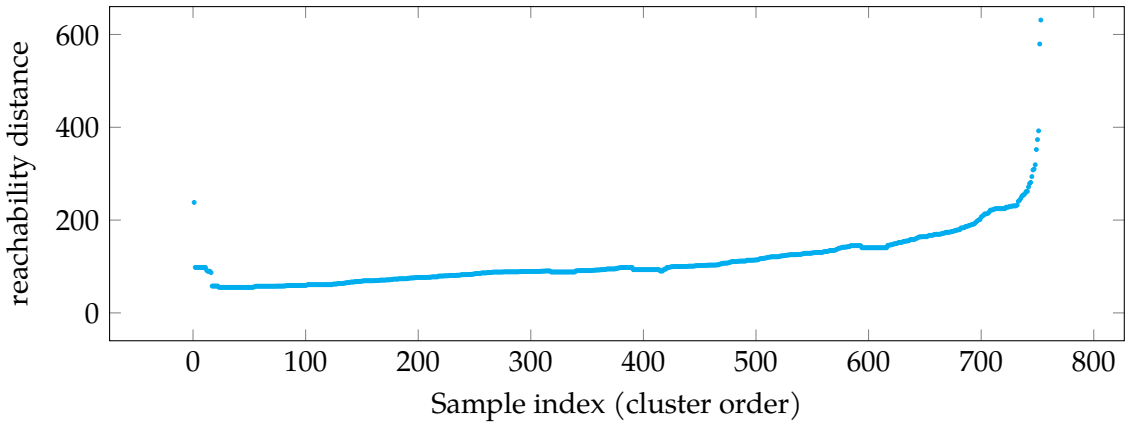


Figure 7.2: Reachability plot for the DTW distance measure on the synthetic data set

7.2.4 Next Steps

As noted in previous sections, none of the attempted approaches produced usable results. Due to time constraints, further analysis using new signal combinations or different approaches was not possible. Perhaps further investigations into this area could produce new insights and result in a working approach for identifying situations based on the event records. This section presents some starting points for further investigations.

Initially, the cluster assignment produced by the prototype was based on the presence of signals instead of the similarity between event records. This was due to placeholders with a large values, which introduced large distances between events that drowned out smaller differences from different signal traces.

Different clustering algorithms may produce better results. For example, Hierarchical Agglomerative Clustering (HAC) has been used previously for clustering in a similar data analysis task with good results.

Different data sets could also be used to improve data analysis. For example, resampling currently uses linear interpolation. By analyzing the relationship between signal traces and stored event records in simulation data, resampling and interpolation could be improved.

Analysis using the synthetic data set pointed at another problem. With more event records included in a data set, differences between an event record and the next most similar might become smaller. When this happens, clustering approaches might start merging clusters, until only a single cluster remains. A hybrid approach could help with this. An initial cluster assignment could be computed on a small data set. This cluster assignment could then be used as a training data set for a classifier, such as a nearest neighbor classifier. A similar pseudo-labelling approach has been employed by Jurczyńska [Jur19] with promising results. Additionally, human verification of a cluster assignment becomes more feasible with a smaller data set.

For most ABA event records, signal traces follow a common trend, which could complicate data analysis. If these common trends can be identified in a data analysis, a preprocessing step can remove trends from event records prior to clustering. Without common trends, differences between individual event records would become more pronounced. This, in turn, could improve the clustering performance.

7.3 Threats to Validity

This section discusses possible threats to validity of my study. It also lists measures that were taken to mitigate these threats.

The first concern is that the research process was performed by a single researcher, introducing the risk of biasing the results. To prevent potential bias stemming from this, the research process was developed in collaboration with my supervisor. In the literature review, selection criteria were kept broad to limit selection bias further. In the interviews, approval of the transcripts served the same purpose.

In the literature review, search was performed using a single search engine. This could have caused important research to be missed. On the other hand, Google Scholar indexes a large amount of research literature, so the chances of missing relevant literature are rather slim.

The number of papers in the literature review was also limited, which could also cause important research to be missed. To compensate for this, a more thorough analysis of the selected papers was performed. Additionally, the results from the literature review were compared against the results from the interviews, which helps reduce bias resulting from the limited number of selected papers.

Additionally, a thorough quality appraisal of the research papers considered in the literature review was skipped. To ensure the quality of the considered research papers, only studies that underwent peer review prior to publication were considered.

In the interviews, participants' responses might have been collected inaccurately or incompletely. The interview audio was recorded to ensure complete collection of participants' responses. In addition, the interview transcripts were approved by the participants to give them a chance to correct statements.

During analysis of the interview transcripts, statements of the participants might have been misinterpreted. To mitigate this issue, the context of statements was carefully considered in the interpretation. Additionally, amending the interpretation of a code, the new interpretation was checked against all uses of the code in the interview transcripts.

Furthermore, an interviewee might have misrepresented information in the interview, for example to avoid constructing a negative image of themselves. In order to ease potential doubts, interviewees were assured that the purpose of the study was not to judge individual employees. The interviews were also conducted anonymously and it was ensured that individual statements cannot be traced back to participants.

8 Conclusion and Outlook

This chapter gives a summary of the thesis and outlines possible directions for further research based on this thesis.

8.1 Conclusion

In this thesis, challenges that practitioners face when working with data pipelines were investigated, with special attention to challenges arising in the context of automotive big data applications. A literature review was conducted to identify challenges for data pipeline management in general and in the automotive context. The identified challenges were compared to challenges identified through semi-structured interviews with data pipeline stakeholders at Daimler Truck AG. One of the most pressing challenges is compliance with data protection regulations. The literature review showed that this challenge is not limited to the automotive context, but applies to data pipeline management in general. Further challenges, such as expert knowledge isolated in knowledge silos and specialized data formats, were identified as unique to automotive contexts. Additionally, this thesis discussed approaches that could alleviate the most pressing challenges identified in the interviews.

Furthermore, this thesis investigated suitable data analysis approaches for automotive event data. A clustering approach for identifying similar situations from event records of an emergency braking system was proposed. An evaluation of the approach was conducted, which unfortunately showed that the approach was unable to adequately identify situations. Instead, the clustering seemed to be based on the presence of signals in the event records. Based on the initial prototype, alternative analysis approaches were investigated and further analysis approaches, which have not been investigated due to time constraints, were outlined.

8.2 Outlook

Further research into the challenges associated with data pipeline management in other contexts could help identify causes for the challenges faced by practitioners. Additionally, studying the history of successful transformations towards data-driven organizations, for example in companies such as Google or Netflix, could reveal learnings that are transferrable to other organizations currently in the process of this transformation.

Future research could also target approaches to simplify the integration of new automotive data sources into data analysis pipelines. In Section 7.1.2, IDLs and reliance on common data formats were named as two approaches, whose feasibility for this purpose could be studied.

This thesis failed in developing a data analysis for identifying situations from event records of an emergency braking system. Nevertheless, this identification of situations is an important step towards further improvement of ADAS. As such, I believe that further studies into new approaches for identifying situations from these event records should be conducted. Section 7.2.4 may serve as an inspiration for such endeavors.

Bibliography

[ABKS99] M. Ankerst, M. M. Breunig, H.-P. Kriegel, J. Sander. "OPTICS". In: *Proceedings of the 1999 ACM SIGMOD international conference on Management of data - SIGMOD '99*. ACM Press, 1999. DOI: 10.1145/304182.304187 (cit. on pp. 22, 23).

[ADA19] Allgemeiner Deutscher Automobil Club. *Ereignisdatenspeicher*. German. July 11, 2019. URL: https://res.cloudinary.com/adacde/image/upload/v1573032982/ADAC-eV/KOR/Text/PDF/event-data-recorder-adac-sp_aeopne.pdf (visited on 01/05/2021) (cit. on p. 19).

[AHK11] S. Adolph, W. Hall, P. Kruchten. "Using grounded theory to study the experience of software development". In: *Empirical Software Engineering* 16.4 (Jan. 2011), pp. 487–513. DOI: 10.1007/s10664-010-9152-6 (cit. on pp. 29, 30).

[ASA19] ASAM e. V. *Measurement Data Format*. Tech. rep. Sept. 30, 2019 (cit. on p. 54).

[AUT20] Automotive Open System Architecture. *ARXML Serialization Rules*. Tech. rep. Nov. 30, 2020 (cit. on p. 54).

[CIM+01] J. T. Correia, K. A. Iliadis, E. S. McCarron, M. A. Smolej, B. Hastings, C. C. Engineers. "Utilizing data from automotive event data recorders". In: *Proceedings of the Canadian Multidisciplinary Road Safety Conference XII, London Ontario*. 2001 (cit. on p. 18).

[CJ74] T. Caliński, H. JA. "A Dendrite Method for Cluster Analysis". In: *Communications in Statistics - Theory and Methods* 3 (Jan. 1974), pp. 1–27. DOI: 10.1080/03610927408827101 (cit. on p. 32).

[CPS20] B. Cartaxo, G. Pinto, S. Soares. "Rapid Reviews in Software Engineering". In: *Contemporary Empirical Methods in Software Engineering*. Springer International Publishing, Mar. 22, 2020, pp. 357–384. DOI: 10.1007/978-3-030-32489-6_13 (cit. on p. 28).

[Dai19] Daimler AG. *Active Brake Assist 5*. Apr. 30, 2019. URL: https://media.daimler.com/marsMediaSite/pic/en/43207744 (cit. on p. 18).

[Dai20] Daimler Truck AG. *Mercedes-Benz Trucks presents two worldwide innovations in their trucks for more safety on the road*. Sept. 23, 2020. URL: https://media.daimler.com/marsMediaSite/doc/en/47504470 (cit. on p. 17).

[DB79] D. L. Davies, D. W. Bouldin. "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2 (Apr. 1979), pp. 224–227. ISSN: 1939-3539. DOI: 10.1109/TPAMI.1979.4766909 (cit. on p. 32).

[GDPR] European Parliament, Council of the European Union. *Regulation (EU) 2016/679*. May 4, 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679 (cit. on pp. 19, 20).

[GHMT17] B. Golshan, A. Halevy, G. Mihaila, W.-C. Tan. "Data Integration". In: *Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems - PODS '17*. ACM Press, 2017. DOI: 10.1145/3034786.3056124 (cit. on pp. 15, 21).

[GSR19] European Parliament, Council of the European Union. *Regulation (EU) 2019/2144*. Nov. 27, 2019. URL: http://data.europa.eu/eli/reg/2019/2144/oj (cit. on p. 20).

[HA05] S. E. Hove, B. Anda. "Experiences from Conducting Semi-structured Interviews in Empirical Software Engineering Research". In: *11th IEEE International Software Metrics Symposium (METRICS'05)*. Como (Sept. 19–22, 2005). Como: IEEE, Sept. 19–22, 2005, pp. 10–23. ISBN: 0-7695-2371-4. DOI: 10.1109/METRICS.2005.24 (cit. on p. 29).

[HA85] L. Hubert, P. Arabie. "Comparing partitions". In: *Journal of Classification* 2.1 (1985), pp. 193–218. ISSN: 1432-1343. DOI: 10.1007/BF01908075 (cit. on p. 32).

[HLV00] B. Hüsemann, J. Lechtenbörger, G. Vossen. "Conceptual Data Warehouse Design". In: *Proceedings of the Second Intl. Workshop on Design and Management of Data Warehouses, DMDW 2000, Stockholm, Sweden, June 5-6, 2000*. Ed. by M. A. Jeusfeld, H. Shu, M. Staudt, G. Vossen. Vol. 28. CEUR Workshop Proceedings. CEUR-WS.org, June 5–6, 2000, p. 6. URL: http://ceur-ws.org/Vol-28/paper6.pdf (cit. on p. 22).

[HMD17] A. Haroun, A. Mostefaoui, F. Dessables. "Data fusion in automotive applications". In: *Personal and Ubiquitous Computing* 21.3 (Feb. 2017), pp. 443–455. DOI: 10.1007/s00779-017-1008-2 (cit. on p. 15).

[Hor85] R. Horn. *Matrix analysis*. Cambridge Cambridgeshire New York: Cambridge University Press, 1985. ISBN: 9780521386326 (cit. on pp. 31, 42).

[JBJ+14] M. Johanson, S. Belenki, J. Jalminger, M. Fant, M. Gjertz. "Big Automotive Data: Leveraging large volumes of data for knowledge-driven product development". In: *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, Oct. 2014. DOI: 10.1109/bigdata.2014.7004298 (cit. on p. 25).

[JNR+20] P. Jovanovic, S. Nadal, O. Romero, A. Abelló, B. Bilalli. "Quarry: A User-centered Big Data Integration Platform". In: *Information Systems Frontiers* (Apr. 2020). DOI: 10.1007/s10796-020-10001-y (cit. on p. 26).

[Jur19] N. Jurczyńska. "City Safety Event Classification using Machine Learning. A binary classification of a multivariate time series sensor data". MA thesis. University of Gothenburg, Nov. 21, 2019. URL: http://hdl.handle.net/2077/62580 (cit. on pp. 25, 58).

[Kim20] M. Kim. "Software Engineering for Data Analytics". In: *IEEE Software* 37 (4 2020), pp. 36–42. ISSN: 1937-4194. DOI: 10.1109/MS.2020.2985775 (cit. on p. 15).

[LKM+15] A. Luckow, K. Kennedy, F. Manhardt, E. Djerekarov, B. Vorster, A. Apon. "Automotive big data: Applications, workloads and infrastructures". In: Santa Clara, CA. Santa Clara, CA: IEEE, 2015, pp. 1201–1210. ISBN: 978-1-4799-9925-5. DOI: 10.1109/BigData.2015.7363874 (cit. on p. 15).

[MBO20] A. R. Munappy, J. Bosch, H. H. Olsson. "Data Pipeline Management in Practice: Challenges and Opportunities". In: *Product-Focused Software Process Improvement*. Springer International Publishing, 2020, pp. 168–184. DOI: 10.1007/978-3-030-64148-1_11 (cit. on pp. 21, 22, 25).

[MK00] D. L. Moody, M. A. R. Kortink. "From enterprise models to dimensional models: a methodology for data warehouse and data mart design". In: *Proceedings of the Second Intl. Workshop on Design and Management of Data Warehouses, DMDW 2000, Stockholm, Sweden, June 5-6, 2000*. Ed. by M. A. Jeusfeld, H. Shu, M. Staudt, G. Vossen. Vol. 28. CEUR Workshop Proceedings. CEUR-WS.org, June 5–6, 2000, p. 5. URL: http://ceur-ws.org/Vol-28/paper5.pdf (cit. on p. 22).

[MRT18] M. Mohri, A. Rostamizadeh, A. Talwalkar. *Foundations of machine learning*. MIT press, 2018. ISBN: 9780262039406. URL: https://books.google.de/books?id=dWB9DwAAQBAJ (cit. on p. 22).

[MT16] N. Miloslavskaya, A. Tolstoy. "Big Data, Fast Data and Data Lake Concepts". In: *Procedia Computer Science* 88 (2016), pp. 300–305. DOI: 10.1016/j.procs.2016.07.439 (cit. on p. 22).

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830. URL: https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html (cit. on pp. 41, 43).

[PVG+19] D. Pevec, H. Vdovic, I. Gace, M. Sabolic, J. Babic, V. Podobnik. "Distributed Data Platform for Automotive Industry: A Robust Solution for Tackling Big Challenges of Big Data in Transportation Science". In: *2019 15th International Conference on Telecommunications (ConTEL)*. IEEE, July 2019. DOI: 10.1109/contel.2019.8848542 (cit. on p. 26).

[RBOW20] A. Raj, J. Bosch, H. H. Olsson, T. J. Wang. "Modelling Data Pipelines". In: *2020 46th Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. IEEE, Aug. 2020. DOI: 10.1109/seaa51224.2020.00014 (cit. on pp. 21, 25).

[RH07] A. Rosenberg, J. Hirschberg. *V-Measure. A Conditional Entropy-Based External Cluster Evaluation Measure*. June 2007 (cit. on p. 32).

[Rou87] P. J. Rousseeuw. "Silhouettes. A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20 (1987), pp. 53–65. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7 (cit. on p. 32).

[SC78] H. Sakoe, S. Chiba. "Dynamic programming algorithm optimization for spoken word recognition". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 26.1 (Feb. 1978), pp. 43–49. ISSN: 0096-3518. DOI: 10.1109/TASSP.1978.1163055 (cit. on p. 23).

[SS20] I. Sucholutsky, M. Schonlau. "'Less Than One'-Shot Learning: Learning N Classes From M<N Samples". In: (Sept. 17, 2020). arXiv: 2009.08449 [cs.LG] (cit. on p. 23).

[StVG] *Straßenverkehrsgesetz*. Mar. 5, 2003 (cit. on p. 21).

[TFV+20] R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, E. Woods. "Tslearn, A Machine Learning Toolkit for Time Series Data". In: *Journal of Machine Learning Research* 21.118 (2020), pp. 1–6. URL: http://jmlr.org/papers/v21/20-091.html (cit. on p. 43).

[TOG19] The Open Group. *ArchiMate 3.1 Specification*. The Open Group, Nov. 2019. URL: https://pubs.opengroup.org/architecture/archimate3-doc/ (cit. on p. 42).

[UMR+15] S. Ulbrich, T. Menzel, A. Reschka, F. Schuldt, M. Maurer. "Defining and Substantiating the Terms Scene, Situation, and Scenario for Automated Driving". In: *2015 IEEE 18th International Conference on Intelligent Transportation Systems*. Sept. 2015, pp. 982–988. DOI: 10.1109/ITSC.2015.164 (cit. on p. 40).

[VEB09] N. X. Vinh, J. Epps, J. Bailey. "Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?" In: *Proceedings of the 26th Annual International Conference on Machine Learning*. ICML '09. Montreal, Quebec, Canada: Association for Computing Machinery, 2009, pp. 1073–1080. ISBN: 9781605585161. DOI: 10.1145/1553374.1553511 (cit. on p. 32).

[WK20] R. Wu, E. J. Keogh. "FastDTW is approximate and Generally Slower than the Algorithm it Approximates". In: *IEEE Transactions on Knowledge and Data Engineering* (2020), pp. 1–1. DOI: 10.1109/tkde.2020.3033752 (cit. on pp. 23, 24).

[WYKN20] Y. Wang, Q. Yao, J. T. Kwok, L. M. Ni. "Generalizing from a Few Examples. A Survey on Few-Shot Learning". In: *ACM Comput. Surv.* 53.3 (June 2020), pp. 1–34. ISSN: 0360-0300. DOI: 10.1145/3386252 (cit. on p. 23).

[WZTE18] T. Wang, J.-Y. Zhu, A. Torralba, A. A. Efros. "Dataset Distillation". In: (Nov. 27, 2018). arXiv: 1811.10959 [cs.LG] (cit. on p. 23).

[YLCT19] Q. Yang, Y. Liu, T. Chen, Y. Tong. "Federated Machine Learning". In: *ACM Transactions on Intelligent Systems and Technology* 10.2 (Feb. 2019), pp. 1–19. DOI: 10.1145/3298981 (cit. on p. 15).

A List of Studies Considered in Literature Review

A.1 Research Question 1.1

[FHL14] J. Fan, F. Han, H. Liu. "Challenges of Big Data analysis". In: *National Science Review* 1.2 (Feb. 2014), pp. 293–314. DOI: 10.1093/nsr/nwt032 (cit. on pp. 33, 34).

[IS15] M. Iqbal, T. Soomro. "Big Data Analysis: Apache Storm Perspective". In: *International Journal of Computer Trends and Technology* 19 (Jan. 2015), pp. 9–14. DOI: 10.14445/22312803/IJCTT-V19P103 (cit. on pp. 33, 34).

[LJ12] A. Labrinidis, H. V. Jagadish. "Challenges and opportunities with big data". In: *Proceedings of the VLDB Endowment* 5.12 (Aug. 2012), pp. 2032–2033. DOI: 10.14778/2367502.2367572 (cit. on pp. 33, 34).

[TEK+16] I. V. Tetko, O. Engkvist, U. Koch, J.-L. Reymond, H. Chen. "BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry". In: *Molecular Informatics* 35.11-12 (July 2016), pp. 615–621. DOI: 10.1002/minf.201600073 (cit. on pp. 33–35).

A.2 Research Question 1.2

[GJ14] X. Ge, J. Jackson. "The Big Data Application Strategy for Cost Reduction in Automotive Industry". In: *SAE International Journal of Commercial Vehicles* 7.2 (Sept. 2014), pp. 588–598. DOI: 10.4271/2014-01-2410 (cit. on pp. 35, 36).

[JBJ+14] M. Johanson, S. Belenki, J. Jalminger, M. Fant, M. Gjertz. "Big Automotive Data: Leveraging large volumes of data for knowledge-driven product development". In: *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, Oct. 2014. DOI: 10.1109/bigdata.2014.7004298 (cit. on pp. 35, 36).

[PVG+19] D. Pevec, H. Vdovic, I. Gace, M. Sabolic, J. Babic, V. Podobnik. "Distributed Data Platform for Automotive Industry: A Robust Solution for Tackling Big Challenges of Big Data in Transportation Science". In: *2019 15th International Conference on Telecommunications (ConTEL)*. IEEE, July 2019. DOI: 10.1109/contel.2019.8848542 (cit. on pp. 35, 36).

[SAFR18] M. Syafrudin, G. Alfian, N. Fitriyani, J. Rhee. “Performance Analysis of IoT-Based Sensor, Big Data Processing, and Machine Learning Model for Real-Time Monitoring System in Automotive Manufacturing”. In: *Sensors* 18.9 (Sept. 2018), p. 2946. DOI: 10.3390/s18092946 (cit. on p. 35).

B Interview Guide

B.1 Preamble

B.1.1 Purpose of this Study

This study focuses on data integration and analysis pipelines in the automotive context. In order to better understand the challenges that accompany the integration and maintenance of such pipelines, we will be conducting interviews.

The goal of this study is NOT to evaluate or rate individual employees. Likewise, we have NO intention to destroy a company's competitive advantage by disclosing vital information about technical excellence or weaknesses in their development organization.

B.1.2 Confidentiality and Anonymity

Everything said in the interviews will be treated as strictly confidential and will not be disclosed to anyone outside our research team without the participant's permission, not even to his/her manager. Results are aggregated and anonymized before publication. In no way can individual statements be traced back to a company or an employee. After the interview, participants also have the possibility to exclude or redact statements before the analysis. The complete interview transcripts will NOT be published.

B.1.3 Interview Process and Analysis

Interviewees will be notified about the general topic of the study so that they are familiar with it and can gather information beforehand should they decide to do so. Interview time is estimated around 30 minutes. We kindly ask your permission to record the interview for transcription. The written transcript will be sent back to the participant for his/her final approval. This is the last chance to remove or redact statements from the interview. After that, the audio will be destroyed and the potentially redacted and approved transcript will be the sole base for our qualitative content analysis.

B.2 Interview Guide

B.2.1 Background

- Role/Tasks
- Years in role
- Team size

B.2.2 Current Situation

- Do you already perform analyses on the VRDU event data?
  - If no, what is stopping you from doing so?
  - If so, what analyses do you perform?
  - How do you "compute" the result of an analysis?
- Are you working in the data pipeline already?
  - If yes, where are you located in the data pipeline?
  - How long did the initial implementation/onboarding take?
- Do you know which tools exist to perform data analysis?
  - Do you have experience in using these tools?
- Do you know what data is available/exists?
  - Do you know where to find such information?

B.2.3 Challenges

- Are you working in the data pipeline already?
  - If no, what is stopping you from doing so?
  - If yes, which challenges do you face when doing so?
- Do you collaborate with other teams (e.g. to share results)?
  - With how many teams do you interact?
  - How does the collaboration work?
  - What challenges do you face when interacting with other teams?
- How do you handle legal restrictions on the data (e.g. from privacy regulation)?

B.2.4 Requirements

- What questions could an analysis answer to be useful to you?

Declaration

I hereby declare that the work presented in this thesis is entirely my own and that I did not use any other sources and references than the listed ones. I have marked all direct or indirect statements from other sources contained therein as quotations. Neither this work nor significant parts of it were part of another examination procedure. I have not published this work in whole or in part before. The electronic copy is consistent with all submitted copies.

---

place, date, signature