Big Automotive Data Preprocessing: A Three Stages Approach

Amal Tawakuli, Daniel Kaiser, Thomas Engel
firstname.surname@uni.lu
University of Luxembourg

ABSTRACT

The automotive industry generates large datasets of various formats, uncertainties and frequencies. To exploit Automotive Big Data, the data needs to be connected, fused and preprocessed to quality datasets before being used for production and business processes. Data preprocessing tasks are typically expensive, tightly coupled with their intended AI algorithms and are done manually by domain experts. Hence there is a need to automate data preprocessing to seamlessly generate cleaner data. We intend to introduce a generic data preprocessing framework that handles vehicle-to-everything (V2X) data streams and dynamic updates. We intend to decentralize and automate data preprocessing by leveraging edge computing with the objective of progressively improving the quality of the dataflow within edge components (vehicles) and onto the cloud.

KEYWORDS

Big Data, Data Prepocessing, Edge Computing, Connected Vehicles

1 INTRODUCTION

Big Automotive Data powers the digital revolution of the automotive industry. It is the lifeblood of Artificial Intelligence and the enabler of intelligent automotive applications and services such as assisted driving and predictive maintenance [17]. Just like many raw materials, Big Data must undergo several operations to be transformed into ready to consume input, or what we call quality data. Such transformations are not easy and may take up to 80% of available recourses [9, 13] for specific applications within a specific domain let alone generic and interdisciplinary applications. New AI projects are often hindered by data preprocessing prerequisites; the difficulty is manifested in finding the complete data, retrieving it, fusing data from multiple sources and transforming the data into quality input that conforms to predefined requirements. Another problem is that data is preprocessed when AI projects are identified and initiated, which increases preprocessing costs and complexity due to the accumulation of raw data. Real-time data add further complexity with low latency processing requirements.

There is no official definition for data preprocessing but from scoping the literature, one can identify a common understanding that data preprocessing is the operations that precede AI algorithms and help transform raw data into quality input. This step is proven to be necessary in creating good models and predictions [13, 15, 22]. Data preprocessing tasks depend on the data, application and the context. It includes operations such as *data integration*, *normalization*, *format conversion* and *noise reduction* [1, 2, 6, 9, 15, 20].

We propose shifting towards automated and progressive data preprocessing with some operations distributed closer to the data sources. By dividing data preprocessing into stages executed by different entities at different locations, the challenges of maintaining quality data can be incrementally conquered. This can be done by leveraging smart sensors and new vehicle E/E architectures [3, 16]. While our concepts can be applicable for several industrial sectors involving Big Data preprocessing, we focus on V2X communication. Our research revolves around the following questions:

- How to reduce resources required for data preprocessing?
- How to decentralize data preprocessing onto the edge to handle stream data and distribute the computation load?
- How to leverage edge computing to yield privacy protection?

2 RELATED WORK

Many data preprocessing solutions exist. In fact, for every AI project there exists a stage for input preparation. These solutions are specific to the problem in hand and focus on batch data. The preprocessing operations often go through several stages of manual modifications and upgrades to extract input data that would improve models and predictions [4, 6, 9, 14, 15, 17, 22].

Automotive research [7, 8, 10, 25] in Big Data focuses on creating edge telematics systems that transmit data collected from vehicle sensors to central cloud-based systems where data preparation and analytics occurs. Only few address the demand for handling stream data or leverage edge components' computational capacities for in vehicle data preprocessing [8, 10]. With these few solutions data preprocessing is restricted to compression for the purpose of reducing network load and costs.

Most generic preprocessing solutions reside in the cloud [12, 21, 23] and tackle the problem as a holistic event. [11, 19] use edge computing for data preprocessing, however, they are human-centric or have fixed and limited preprocessing operations, see section 3.

3 DATA PREPROCESSING

The scope of data preprocessing can be categorized as follows: 1) data access and fusion, 2) parsing and decoding, 3) partitioning/windowing and schedulling, 4) cleaning and feature extraction, 5) transformation and sampling, 6) augmentation and tagging, 7) encryption and authentication. We will attempt to cover many of these data preprocessing categories emphasising on creating a generic and automated data preprocessing solution. We envision the following challenges to overcome: 1) Identify data preprocessing operations suitable for edge components and those more fitted for backend systems. 2) Identify and categorize preprocessing operations that can be applied on different data from different sources.

CSCS '19, October 08, 2019, Kaiserslautern, Germany

ACM ISBN ??…$15.00
https://doi.org/10.1145/1122445.1122456

Our research does not aim at creating a comprehensive solution that would cover all operations within all the categories of data preprocessing but rather find an optimal mapping between preprocessing operations and the three stages, see section 4, that would gradually create higher quality data. 3) Preprocess data streams and handle their high throughput and low-latency requirements. 4) Develop a software architecture designed for edge components, particularly vehicles. 5) Evaluate the effectiveness of the data preprocessing solution and its impact on AI algorithms.

4 THE THREE STAGES APPROACH

Our data preprocessing solution is a hybrid in the sense that it will run on the cloud and at the edge. To progressively improve data quality, we propose three stages of data preprocessing:

4.1 Smart Sensors

Extending sensors' capabilities beyond data acquisition is an essential step towards fulfilling the high throughput and low latency requirements of stream data [18]. We propose exploiting smart sensors to perform frontline preprocessing tasks as data are collected. Given their limited computation capacities, smart sensors will perform simple preprocessing operations. Due to their close proximity to the data, they can perform data-specific tasks. The preprocessing operations will be deployed as a query plan via the central in-vehicle engine making the solution versatile and compatible with OTA updates. Exploiting smart sensors further distributes the preprocessing work load between the cloud and the edge.

As reprogrammable smart sensors are not widely available on the market, we plan to create our own smart sensors using the single-board computer Raspberry Pi connected to a sensor [24].

4.2 Central in-Vehicle Preprocessing Engine

The second stage adds another preprocessing layer at the edge. The engine handles data from multiple sensors and has greater computation capacity allowing it to perform complex preprocessing operations. As it is closer to data sources, the engine will be designed to handle data in real-time. We aim to provide a powerful in-vehicle stream data preprocessing engine by exploiting parallel computing.

In an automotive scenario, the engine would connect to the vehicle's bus systems to access sensor data. The engine will perform operations from the different data preprocessing categories. Compiling a general and optimal set of preprocessing operations for different datasets is a challenge. We envision that operations such as meta-data tagging (enables data lineage), data compression (reduces data volume) and format conversion can be applied on different datasets and would have optimal impact if deployed on the edge. The engine will also apply preprocessing requirements received form the cloud, which would enable the execution of operations such as feature extraction and instance selection on different datasets. The preprocessed data are either injected into in-vehicle AI algorithms or transmitted to the cloud. This requires communication management to be integrated in the engine to provide certain levels of privacy, see section 5. The engine will be designed to be extensible and resilient to change by receiving OTA updates (new preprocessing requirements). The engine either executes the updates or relays them to the sensors.

4.3 Cloud Data Preprocessing Infrastructure

The final stage consists of a cloud-centric system that performs more complex preprocessing operations given its powerful computation capabilities and available storage. The sensor data received are no longer characterized as real-time data and can either be processed as it arrives or stored for batch processing. For such reasons, deep learning algorithms for deriving optimal training sets and detecting noise/outlier could be ideal for this stage.

In our framework, the data have been preprocessed in previous stages, which we expect to reduce resource consumption and costs incurred in cloud systems to prepare data. The cloud system will also manage and communicate new or modified preprocessing operations/requirements to the edge (OTA updates). Our project will not cover data storage and will focus mostly on the edge stages.

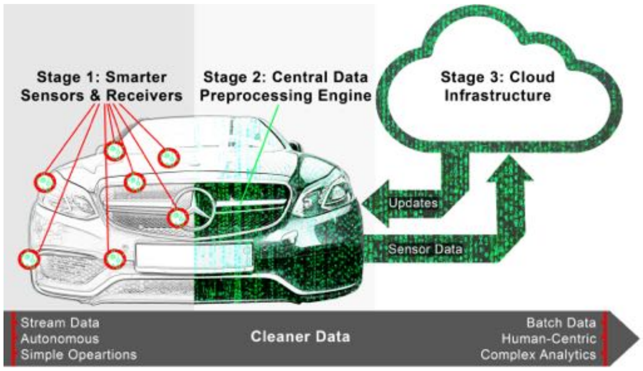

Figure 1: The three stages and their effects on data quality

5 SECURITY AND PRIVACY PROTECTION

Adding security and privacy protection typically comes at a performance cost, however, shifting preprocessing to the edge will not only allow us to increase efficiency but also improve privacy protection since a significant amount of data will not leave the edge. Besides this innate data protection, we will consider security and privacy while designing our framework. We intend to protect the data via the following approaches: 1) Anonymizing (e.g. by randomization and noise injection) transmitted data, which is necessary for GDPR [5] compliance. 2) Authenticating transmitted data, which allows for veracity and, in turn, provable data lineage.

6 CONCLUSION

Big data is a new source of value for many industries including the automotive industry, however, many opportunities to capitalize data is hindered by expensive and complex preprocessing tasks, which could break or make the analytics phase that follows. We propose a new approach for tackling the challenges of preprocessing data by deriving a generic framework that automates and distributes the tasks to create a dataflow with gradually improving data quality. We intend to conduct this applied research in collaboration with an industry partner. Given the nature and objectives of the proposed thesis, we aim to develop prototypes that handle real sensor data and ideally perform tests on edge (vehicle) components.

REFERENCES

[1] Google Inc 2019. *Data preprocessing for machine learning: options and recommendations*. Google Inc. https://cloud.google.com/solutions/machine-learning/data-preprocessing-for-ml-with-tf-transform-pt1

[2] Stamatios-Aggelos N. Alexandropoulos, Sotiris B. Kotsiantis, and Michael N. Vrahatis. 2019. Data preprocessing in predictive data mining. *The Knowledge Engineering Review* 34 (2019). https://doi.org/10.1017/s026988891800036x

[3] Ondrej Burkacky, Johannes Deichmann, Georg Doll, and Christian Knochenhauer. 2018. *Rethinking car software and electronics architecture*. McKinsey & Company. https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/rethinking-car-software-and-electronics-architecture

[4] Matthias Auf der Mauer, Tristan Behrens, Mahdi Derakhshanmanesh, Christopher Hansen, and Stefan Muderack. 2018. Applying Sound-Based Analysis at Porsche Production: Towards Predictive Maintenance of Production Machines Using Deep Learning and Internet-of-Things Technology. (sep 2018), 79–97. https://doi.org/10.1007/978-3-319-95273-4_5

[5] EU. 2016. REGULATION (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). *Official Journal of the European Union* Volume 59 (2016), 51–53. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L_.2016.119.01.0001.01.ENG&toc=OJ:L:2016:119:TOC

[6] Salvador Garcia, Julian Luengo, and Francisco Herrera. 2015. *Data Preprocessing in Data Mining*. Springer International Publishing. https://doi.org/10.1007/978-3-319-10247-4

[7] Amir Haroun, Ahmed Mostefaoui, and Francois Dessables. 2017. A Big Data Architecture for Automotive Applications: PSA Group Deployment Experience. In *2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)*. IEEE. https://doi.org/10.1109/ccgrid.2017.107

[8] Bastian Havers, Romaric Duvignau, Hannaneh Najdataei, Vincenzo Gulisano, Ashok Chaitanya Koppisetty, and Marina Papatriantafilou. 2019. DRIVEN: a Framework for Efficient Data Retrieval and Clustering in Vehicular Networks. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE. https://doi.org/10.1109/icde.2019.00201

[9] Yu Huang, Mostafa Milani, and Fei Chiang. 2018. PACAS: Privacy-Aware, Data Cleaning-as-a-Service. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE. https://doi.org/10.1109/bigdata.2018.8622249

[10] Mathias Johanson, Stanislav Belenki, Jonas Jalminger, Magnus Fant, and Mats Gjertz. 2014. Big Automotive Data: Leveraging large volumes of data for knowledge-driven product development. In *2014 IEEE International Conference on Big Data (Big Data)*. IEEE. https://doi.org/10.1109/bigdata.2014.7004298

[11] Alpa Kohli, Nick Schonning, and JiayueHu. 2019. *What is Azure Data Box Edge?* Microsoft. https://docs.microsoft.com/en-us/azure/databox-online/data-box-edge-overview

[12] Sanjay Krishnan, Michael J. Franklin, Ken Goldberg, Jiannan Wang, and Eugene Wu. 2016. ActiveClean:An Interactive Data Cleaning Framework For Modern Machine Learning. In *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16*. ACM Press. https://doi.org/10.1145/2882903.2899409

[13] Sanjay Krishnan and Eugene Wu. 2019. AlphaClean: Automatic Generation of Data Cleaning Pipelines. *arXiv e-prints*, Article arXiv:1904.11827 (Apr 2019), arXiv:1904.11827 pages. arXiv:cs.DB/1904.11827

[14] Jacob Langner, Johannes Bach, Lennart Ries, Stefan Otten, Marc Holzapfel, and Eric Sax. 2018. Estimating the Uniqueness of Test Scenarios derived from Recorded Real-World-Driving-Data using Autoencoders. In *2018 IEEE Intelligent Vehicles Symposium (IV)*. IEEE. https://doi.org/10.1109/ivs.2018.8500464

[15] Canchen Li. 2019. Preprocessing Methods and Pipelines of Data Mining: An Overview. *arXiv e-prints*, Article arXiv:1906.08510 (Jun 2019), arXiv:1906.08510 pages. arXiv:cs.LG/1906.08510

[16] Shaoshan Liu, Liangkai Liu, Jie Tang, Bo Yu, Yifan Wang, and Weisong Shi. 2019. Edge Computing for Autonomous Driving: Opportunities and Challenges. *Proc. IEEE* 107, 8 (aug 2019), 1697–1716. https://doi.org/10.1109/jproc.2019.2915983

[17] Andre Luckow, Ken Kennedy, Fabian Manhardt, Emil Djerekarov, Bennie Vorster, and Amy Apon. 2015. Automotive big data: Applications, workloads and infrastructures. In *2015 IEEE International Conference on Big Data (Big Data)*. IEEE. https://doi.org/10.1109/bigdata.2015.7363874

[18] Andre Luckow, Ken Kennedy, Marcin Ziolkowski, Emil Djerekarov, Matthew Cook, Edward Duffy, Michael Schleiss, Bennie Vorster, Edwin Weill, Ankit Kulshrestha, and Melissa C Smith. 2018. Artificial Intelligence and Deep Learning Applications for Automotive Manufacturing. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE. https://doi.org/10.1109/bigdata.2018.8622357

[19] Behshad Mohebali, Amirhessam Tahmassebi, Amir H. Gandomi, and Anke Meyer-Baese. 2019. A big data inspired preprocessing scheme for bandwidth use optimization in smart cities applications using Raspberry Pi. In *Big Data: Learning, Analytics, and Applications*, Fauzia Ahmad (Ed.). SPIE. https://doi.org/10.1117/12.2517440

[20] Sana Mushtaq. 2019. *Data preprocessing in detail*. IBM. https://developer.ibm.com/articles/data-preprocessing-in-detail/

[21] Alexis Perrier, Kishore Ayyadevara, and Giuseppe Ciaburro. 2018. *Hands-On Machine Learning on Google Cloud Platform*. PACKT PUB. 97–126 pages. https://www.ebook.de/de/product/33035072/alexis_perrier_kishore_ayyadevara_giuseppe_ciaburro_hands_on_machine_learning_on_google_cloud_platform.html

[22] Sijie Ruan, Ruiyuan Li, Jie Bao, Tianfu He, and Yu Zheng. 2018. CloudTP: A Cloud-Based Flexible Trajectory Preprocessing Framework. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE. https://doi.org/10.1109/icde.2018.00186

[23] Dawid Rutowicz. [n.d.]. *Clean Frames*. https://github.com/funkyminds/cleanframes

[24] Naveen Kumar Singa, Nilesh Jadhav, and Bony Mathew. 2018. Distributed computing using SMART Sensors in Industrial automation framework. In *2018 Second International Conference on Green Computing and Internet of Things (ICGCIoT)*. IEEE. https://doi.org/10.1109/icgciot.2018.8753076

[25] Mingming Zhang, Tianyu Wo, Tao Xie, Xuelian Lin, and Yaxiao Liu. 2017. CarStream. *Proceedings of the VLDB Endowment* 10, 12 (aug 2017), 1766–1777. https://doi.org/10.14778/3137765.3137781