Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 204 (2022) 817–824

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Industry Sciences and Computer Sciences Innovation

Big Data Analytics for Vehicle Multisensory Anomalies Detection

Ana Xavier Fernandes[a*], Pedro Guimarães[a,b], Maribel Yasmina Santos[a]

[a]*ALGORITMI Research Center, Dep. Information Systems, University of Minho, Guimarães, Portugal*
[b]*EPMQ – IT. CCG ZGDV Institute, Guimarães, Portugal*

Abstract

Autonomous driving is assisted by different sensors, each providing information about certain parameters. What we are looking for is an integrated perspective of all these parameters to drive us into better decisions. To achieve this goal, a system that can handle these Big Data issues regarding volume, velocity and variety is needed. This paper aims to design and develop a real-time Big Data Warehouse repository, integrating the data generated by the multiple sensors developed in the context of IVS (In-Vehicle Sensing) systems; the data to be stored in this repository should be merged, which will imply its processing, consolidation and preparation for the analytical mechanisms that will be required. This multisensory fusion is important because it allows the integration of different perspectives in terms of sensor data, since they complement each other. Therefore, it can enrich the entire analysis process at the decision-making level, for instance, understanding what is going on inside the cockpit.

Keywords: Big Data Warehouse; Big Data; ETL;

1. Introduction

Mobility Service Providers (MSPs), such as Waymo and Lyft [1], who own fleets of autonomous and shared vehicles that provide mobility services, like robot taxis and ride sharing, have identified that an understanding of abnormal events occurring inside the vehicle is a valuable insight for fleet management. Understanding abnormal events enables MSPs to define possible actions that aim to ensure occupants' safety and comfort [2]. Examples of events or activities that may affect occupants' comfort may relate to the occurrence of bad smells, resulting from spoiled food, human vomit or smoking, among others. In this view, an activity is some action performed by an occupant that lasts a considerable amount of time and an event is something that occurs inside the vehicle for a short period of time. Regarding safety, occupants arguing or physically fighting inside the vehicle can result in harm to

* Ana Xavier Fernandes. Tel.: +351968199954.
E-mail address: ana.xavier.fernandes@gmail.com

10.1016/j.procs.2022.08.099

occupants. These events or activities that are detected or even prevented by drivers are now more likely to happen when no human supervision exists [3]. The range of activities or events happening inside vehicles is substantial. Normal activities are perceived as occupant behaviours occurring inside the vehicle that MSPs consider not to have a negative impact on the occupant's sense of comfort and safety. As an example of these activities, generally accepted as normal, are occupants speaking to each other, using radio or texting using mobile phones. This type of activity can occur in moving or stopped vehicles [3]. The work here described was developed in the project *Detection of anomalies from fusion of multiple sensors* and is part of the defined vision for the Intelligent Cockpit, aiming to develop solutions to enable anomaly detection based on the fusion of data from several sensors that monitor the vehicle interior (In-Vehicle Sensing – IVS). A major advantage of sensory fusion systems is the capability of crossing information generated by several sensors. Data coming from several sensors describe normal processes or behaviours occurring inside the vehicle. Anomaly detection aims to understand when these systems behave in abnormal states by considering single or combined sources of sensory data. In this sense, this paper aims to design and develop a real-time Big Data Warehouse repository that will store unstructured data (such as audio data) originated from sensors installed in the interior of a vehicle. This data can be integrated and processed in a distributed way using Big Data concepts and technologies, hence being aligned with the Industry 4.0 agenda of the multinational automobile partner company [4], [5]. The task of storing unstructured data in a BDW is an important one, since it allows the data to be used in future queries and allows the analysis of performance indicators related to such data. Also, the BDW should not only be used to store the unstructured data obtained from the sensors, but should also be used to store relevant metadata, i.e., other data that can add additional context to the unstructured data that is being stored. On a different perspective, the BDW can also be used to store other data resulting from the execution of the Machine Learning algorithms, including predictive attributes in the BDW. This paper is organized as follows: Section 2 summarizes related work in this field. Section 3 presents the proposed Big Data Architecture that includes the BDW for multisensory fusion. Section 0 describes the analytical environment in which analytical dashboards allow the analysis of the integrated data. Section 5 concludes with some remarks and guidelines for future work.

2. Related Work

Kumar et al. proposed a model that enables the recognition, collection, storage and analysis of data collected from vehicles. This model allows collecting all the information from the vehicles through Flume, transmitting it to HDFS (Hadoop Distributed File Systems) through Kafka, and analyzing and processing it using Spark. This work proposes a model in which the Hadoop ecosystem can be applied to the connected car environment, but has the disadvantage of storing in HDFS all the messages collected from the vehicles without classification. The work of Han et al. collects vehicle interior data and passenger's body reaction information through OBD and smart watches. The collection and processing phases were designed using GS1, a global standard for business communication, to establish standards for sharing information collected and stored in a connected environment with the outside world. However, the platform proposed in the paper focuses on collecting vehicle information and has the disadvantage that the messages shared among connected units are limited to simple vehicle status information such as driving and stopping times. Sung et al. intended to develop a driving environment prediction platform based on road traffic information collected for driver safety. Here the method is changed to collect data in real-time from mobile-type sensors rather than from existing fixed-type sensors, thus allowing the collection of several big data sets. The data is analyzed on a driving environment analysis platform based on the Hadoop ecosystem, and information is provided through a web environment. Yoo et al. proposed a platform that includes a big data processing system that collects and processes vehicle big data generated from connected cars and a messaging system that delivers large-capacity vehicle sensor data and traffic information in real time. The platform analyzes and processes the loaded data using Spark and visualizes the results with Zeppelin. The data processed in the platform transmits a message to the central LDM through Kafka for vehicle services. However, as big data increases, it is difficult to obtain the location and type of the existing data, so it is crucial to optimize the data management. Guerreiro et al. proposed an ETL (extract, transform and load) architecture for intelligent transportation systems, addressing an application scenario on dynamic toll charging for highways. The proposed architecture is capable of handling real-time and historical data using Big Data technologies such as Spark on Hadoop and MongoDB. The proposal in this paper stands out from the others by collecting, storing and managing data from multiple sensors of different natures, such as audio, gas and air quality that when combined offer

differentiated multisensorial analytical value and provide enhanced and more complete analyses regarding the vehicle cockpit. For instance, this would provide useful insights to car fleet managers by tracking multiple events during the use of the vehicle such as people coughing, yelling or smoking, humidity and temperature levels and air emissions inside the car. Thus, we propose a Hadoop big data scalable platform that includes a big data processing system using Spark, that collects and processes vehicle big data providing valuable insights through an interactive PowerBI dashboard.

3. Proposed Architecture for the Big Data Architecture

The proposed architecture, Figure 1, empowers the integration of data generated by multiple sensors developed in the context of IVS systems. The data to be stored in this architecture should be combined, which will imply its processing, consolidation and preparation to be compliant with the domain analytical requirements. This will be demonstrated in the following sections. This architecture is structured by three main components: Data Sources, Big Data Cluster and Visualization Tools. The Data Sources can be of different natures, depending on the data domains they produce/handle: data can be structured, semi-structured, or non-structured. More than this, the data sources may have data that is produced at different speeds, with different sizes and formats, thus justifying the need of a Big Data approach [11]. This will be more detailed in section 3.1. The Big Data Cluster component integrates two areas, which are the Data Lake and the Big Data Warehouse. The Data Lake is used to support the storage of any kind of data and processes such as Raw Data and Data Pipelines. The Big Data Warehouse stores data modelled as Analytical Objects, representing highly independent and autonomous entities with a focus on an analytical subject in terms of decision support [12]. This will be further detailed in sections 3.1 and 3.3, respectively. Once all the data has been integrated and properly stored into the Big Data Warehouse, it is possible to analyze it in the Visualization Tools component. This includes data visualizations that support analytical decision making based on the modelled Analytical Objects. In this paper, three Analytical Objects are addressed, namely the ones that integrate data from three different sensors in the vehicle: Microphone, BME (gas sensor) and Particle.

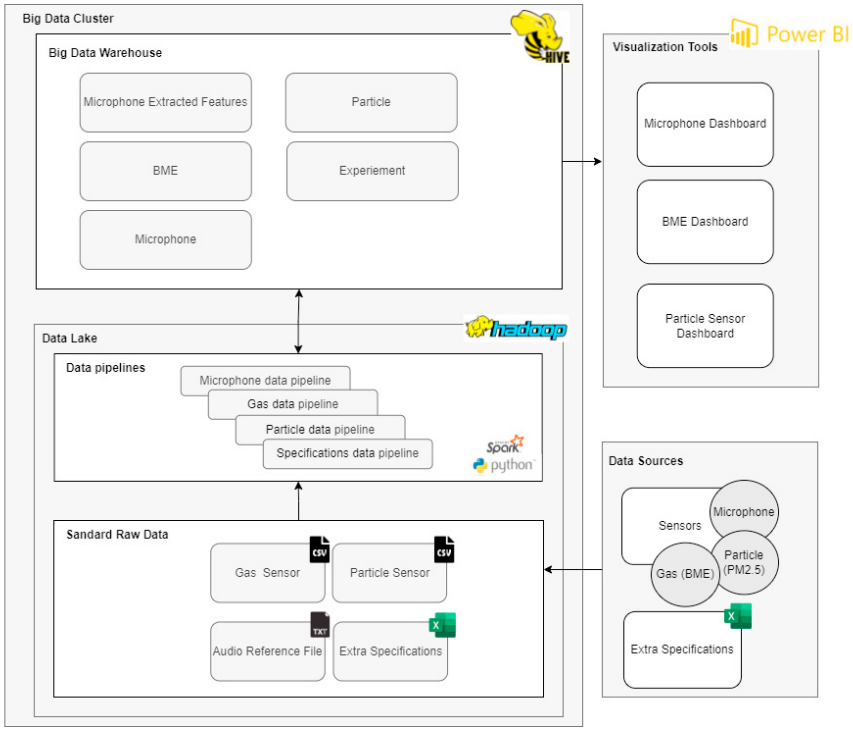

Fig. 1. Proposed Architecture for the Big Data Architecture

3.1. Data Acquisition and Data Lake

The data acquisition process carried out and later shared with the development team covers a multiple array sensor data, including microphone, particle sensor and gas sensor. This sample data was used to design and build the Data Lake and is mainly divided into two categories. A stationary labelled data, which means that data was collected when the vehicle was not moving, and moving data acquisition, when the vehicle was in fact moving on the road. This data acquisition was led by a series of different vehicles ranging from internal combustion engines (petrol, diesel, LPG) to a BEV (Battery Electric Vehicle), under different environment conditions and some other conditional factors such as open or closed windows and radio sound. In order to collect, store and scale the data that is being generated by the multiple car sensors, we decided to define a data storage structure strategy based on features such as data labelling, date, acquisition session (run) or data type.

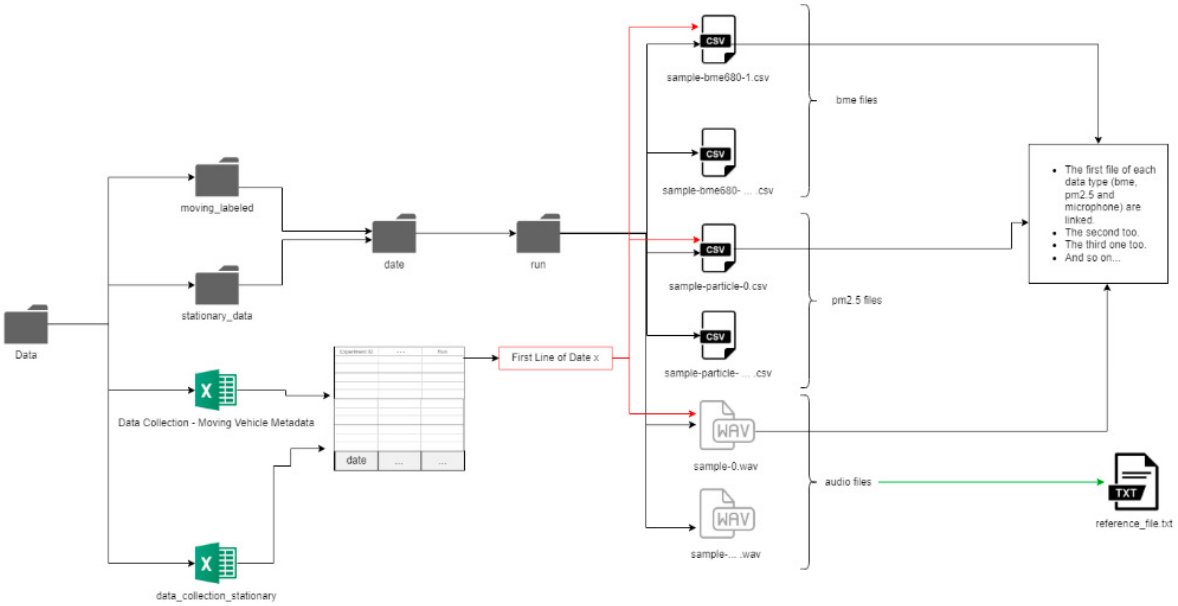

Fig. 2. Data Lake

There are three types of data formats: structured data, semi-structured data and unstructured data. Structured data is made of clearly defined data types, while unstructured data is made of data that is generally not easily searchable. Semi-structured data is in the middle of these two types [13]. The Figure 2 describes the Data Lake structure model developed under Apache HDFS technology on a multi node configuration cluster. Regarding our demonstration case, we covered all types of data formats in this project: unstructured data, using WAV files and TXT files, semi-structured data, using CSV files, and structured data, using Excel files. After the acquisition stage, data is divided and stored into two main folders, each one being the acquisition data type (moving or stationary). From here, we have defined a hierarchical structure of files and folders, so that, on each of these folders, there are sub folders that store the date when the data was collected. Furthermore, there are also folders that organize the data by the identifier of this collection (called run). Each run folder contains three types of files collected from each device: microphone, particle sensor (PM2.5) and gas sensor (BME). The particle sensor provides multiple attributes related to air quality sensing. The entries observed in the CSV files were the pm10 standard, pm25 standard, pm100 standard, pm10 env, pm25 env, pm100 env, particles 03um, particles 05um, particles 10um, particles 25um, particles 50um and particles 100um. The gas sensor measures relative humidity, barometric pressure, ambient temperature, altitude and gas. The CSV files contain an entry for each measure. At last, Excel files contain a wide number of specifications for each data collection event helping in a better understanding of the data. As said before, all data is linked, but first it is necessary to understand two things:

- The files generated in each run are connected according to their iteration number, i.e., for instance, as shown in Figure 2, the “sample-0.wav” file, the “sample-bme680-1.csv” file and the “sample-particle-0.csv” file, despite being from different sensors, were all collected at the same run and are therefore connected;
- Each run is composed of several sheets labelled with the date of collection. In each of the sheets, as mentioned above, there are several features, such as the "run" identifier.

Using this approach, by setting data partitions, we can assure scalability of data taking into account the regular and specific multinational automobile company´s data source formats.

3.2. Data Profiling and Quality

The analysis of data quality in a Big Data context is often a complex task, due to the challenges imposed to the various existing technologies in analyzing data with such characteristics (i.e., high volume, variety, and velocity). On the other hand, this is also a considerably time-consuming task, which contributes to the complexity of performing this task. In this sense, to reduce the complexity of the analysis of data quality, as well as to reduce the time spent on this analysis, a script using python was developed to automate this process [11, 14]. Once connected to the Big Data cluster (technological infrastructure of the partner company named RB Analytics), we designed the data quality pipelines and verified the quality of the data. After fetching all data from HDFS, using pyspark, we developed four data frames: one for all BME files, one for all PM2.5 files, one for all audio files and the last one with the data collection specifications. Later, we converted these pyspark dataframes into pandas dataframes, so that we could use some data quality libraries, such as pandas-profiling or sweatvizz. This pandas library contains functions that help us to better understand data. There is a set of statistical functions to check the quality of the data, for instance: data type, row count, missing values, unique values, "top" and "freq" measures as stated in Table 1.

Table 1. Description of the parameters automatically generated for the data quality analysis

Data Quality Analysis	Parameters Meaning
data type	Data type of each column
row count	Count of the number of rows in each column
missing values	Count of the number of missing values in each column
unique values	Count of the number of unique values in each column
top	The most common value
freq	The most common value’s frequency

For each file, a general analysis was carried out and contains all the features described in previous sections. As a validation process, this data quality analysis had also been previously performed in a manual way by sampling data, in order to compare both approaches (manual and scripted), so that we could ensure that the script is correctly outputting the analysis. Taking this into account, after executing the script, Table 2 shows an example of the generated analysis as described before.

Table 2. Data analysis of BME files (from gas sensor)

	data type	row count	missing values	unique values	top	freq
temperature	object	39640	0	3158	17.99	176
gas_resistence_in_ohms	object	39640	0	4589	150917.00	77
humidity	object	39640	0	4904	31.03	44
pressure	object	39640	0	5892	990.00	92
altitude	object	39640	0	22700	195.38	21

Regarding the gas sensor (BME), the dataset provided is very complete, without missing values with a high cardinality in the altitude attribute. The data quality pipelines corrected some errors detected during the data quality

analysis phase, such as data with values equal to null. These transformations were also important to structure the data according to the analytical requirements and to later load it into the Big Data Warehouse. After the data is aligned with the data quality requirements, it is loaded into Hive (the data warehouse storage system available in Hadoop). Afterwards, PowerBI will fetch this data in order to provide the information in a more intuitive way heading towards adding value to decision making.

3.3. Big Data Warehouse

This subsection describes the main steps followed for the development of the BDW. In this task, the approach described in [15] was considered and adapted for this particular case. The adopted methodology includes the following major phases:

- Identify the application domain conceptual model (with an ER (Entity-Relationship) diagram, for instance);
- Classify the domain entities according to their analytical and descriptive value;
- Identify the BDW conceptual model (based on the Analytical Objects that include analytical and descriptive attributes).

Considering all the available data, the ER diagram was identified in order to understand the relationships between all existing data. With the ER diagram, a matrix that maps the entities and their descriptive and analytical value was elaborated. This is a matrix that helps in the modelling approach as an intermediate step between the ER diagram and the BDW conceptual model. This was developed based on the set of design rules and patterns presented and explained in [15]. The final conceptual model, as illustrated in Figure 3, makes use of the application domain knowledge present in the ER and also the association between the concepts expressed in the matrix. Summarizing, the data modelling approach includes data modelling rules and patterns, practitioners from the domain conceptual model to a BDW conceptual model. This conceptual model consists in four Analytical Objects (AO) - "Microphone", "Microphone_Extracted_Features", "BME" and "Particle" -, a Complementary Analytical Object (CAO) - "Experiment" - and a Special Object (SO) - "Date". Each AO, except "Microphone_Extracted_Features", has attributes related to each feature extracted from each of the sensors. The "Microphone_Extracted_Features" table contains information about total of anomalies and total non-anomalies by date, run and audio file, and has an extra attribute ("predicted") that identifies whether this data is "predicted" (through machine learning inputs) or "labelled", with the value "yes" corresponding to "predicted" data and "no" corresponding to "labelled" data. The "Experiment" table contains all the data that are shared by the microphone, BME and Particle analytical objects. Lastly, the "Date" table contains the temporal view of the BDW, i.e. when the data was generated.

4. Analytical Environment

After the integration and consolidation for all the data in the BDW, and in order to take advantage of this data for decision support activities, a set of analytical dashboards were designed and built. All the dashboards were validated by the domain expert users at the partner company. During the development phase, the PowerBI platform was used in order to build mockups to support this analytical analysis. In this sense, the dashboards were designed aligned with the main sensor source data, i.e. microphone, gas sensor (BME680) and particle sensor (PM2.5). The microphone dashboard, Figure 4 contains several filters in order to facilitate the user's analysis of the data. By default, it displays an overall integrated view of the available data for that sensor. However, using the available filters and selecting a specific value for a given attribute, the dashboard is updated dynamically. The user can view the total number of anomalies per date, and these can be "predicted" ("yes") or "labelled" ("no"). Also, the total number of anomalies for each car used in the data collection is available, showing information about the car itself, brand and model, and the corresponding id. Finally, it is also possible to visualize the total number of anomalies for each event and the status of the wipers, air conditioning, windows and radio, and a table containing more detailed

information about extra specifications, such as the average speed of the car, the location of the car at the time of data collection, whether the windows were open or closed, the type of road, the type of fuel of the car, etc.

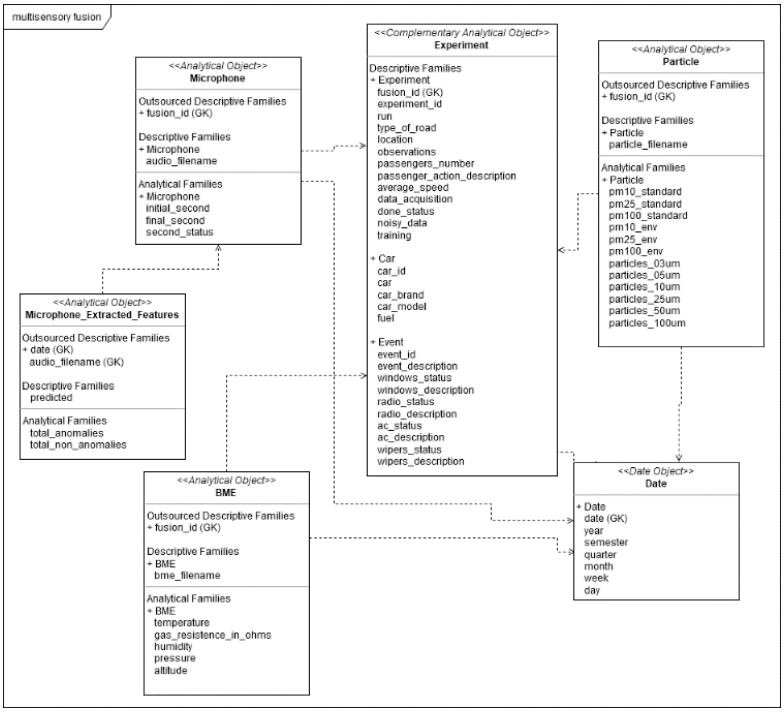

Fig. 3. Big Data Warehouse

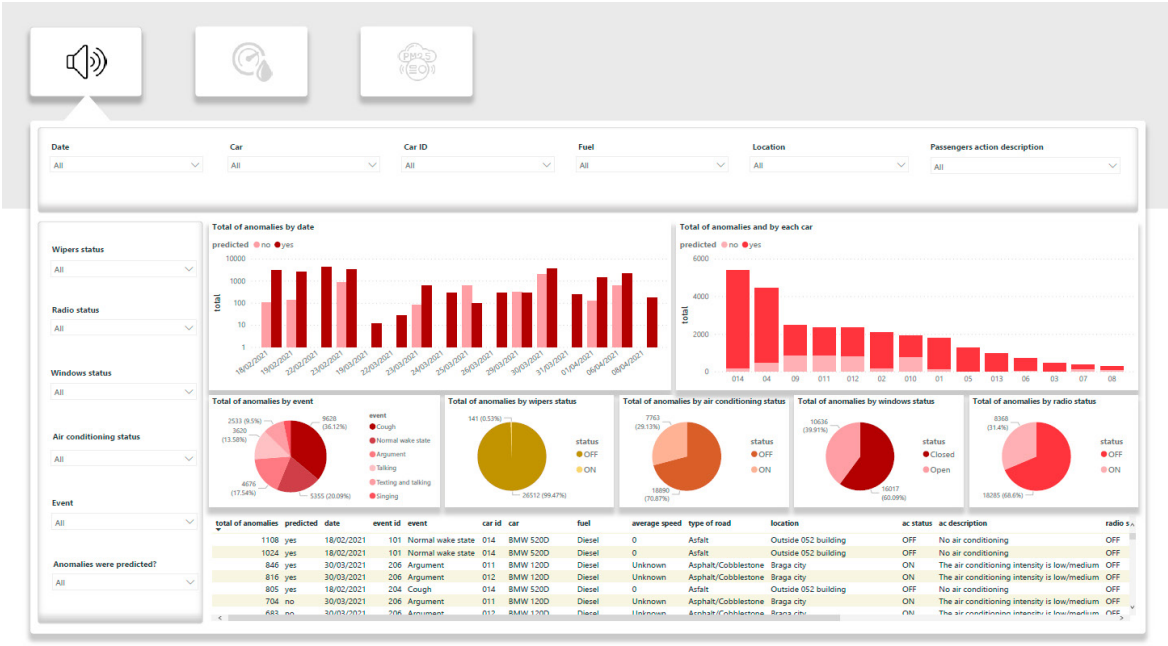

Fig. 4. Microphone Dashboard

5. Conclusion

This paper proposed a Big Data system that includes a BDW as its main analytical storage component to manage large amounts of data generated by multiple sensors inside a vehicle. All the design and development are considered as a technological environment in the Hadoop ecosystem. The available data was processed and stored in order to extract analytical value from it. The ETL process was used as the basis, where data was extracted manually since, due to the partner company's cluster constraints, there was no way to extract the data directly from the source. The BDW is used as the source of integrated data for the analytical environment made available and that, at this moment, includes three dashboards, one for each of the sensors (BME, PM2.5 and microphone).

Acknowledgements

This work has been supported by FCT – *Fundação para a Ciência e Tecnologia* within the R&D Units Project Scope: UIDB/00319/2020 and by the European Structural and Investment Funds in the FEDER component, through the Operational Competitiveness and Internationalization Programme (COMPETE 2020) [Project nº 039334; Funding Reference: POCI-01-0247-FEDER-039334].

References

[1] Kim, S. et al., *Autonomous taxi service design and user experience.* International Journal of Human–Computer Interaction, 2020. **36**(5): p. 429-448.

[2] Pereira, P.J. et al., *Using deep autoencoders for in-vehicle audio anomaly detection.* Procedia Computer Science, 2021. **192**: p. 298-307.

[3] Joshi, A. et al. *Protocol for Eliciting Driver Frustration in an In-vehicle Environment*. in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. 2019. IEEE.

[4] Santos, M.Y. et al., *A big data system supporting bosch braga industry 4.0 strategy.* International Journal of Information Management, 2017. **37**(6): p. 750-760.

[5] Costa, E. et al., *Evaluating partitioning and bucketing strategies for Hive-based Big Data Warehousing systems.* Journal of Big Data, 2019. **6**(1): p. 34.

[6] Kumar, S. et al. *Changing the world of autonomous vehicles using cloud and big data*. in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*. 2018. IEEE.

[7] Han, J. et al. *GS1 Connected Car: An Integrated Vehicle Information Platform and Its Ecosystem for Connected Car Services based on GS1 Standards*. in *2018 IEEE Intelligent Vehicles Symposium (IV)*. 2018. IEEE.

[8] Sung, H.K. et al. *Development of road traffic analysis platform using big data*. in *International Conference on Advances in Big Data Analytics*. 2017.

[9] Yoo, A. et al., *Implementation of a sensor big data processing system for autonomous vehicles in the C-ITS environment.* Applied Sciences, 2020. **10**(21): p. 7858.

[10] Guerreiro, G. et al. *An architecture for big data processing on intelligent transportation systems. An application scenario on highway traffic flows*. in *2016 IEEE 8th International Conference on Intelligent Systems (IS)*. 2016. IEEE.

[11] Cai, L. et al., *The challenges of data quality and data quality assessment in the big data era.* Data science journal, 2015. **14**.

[12] Santos, M.Y. et al., *Big Data: Concepts, Warehousing, and Analytics*. 2020: River Publishers.

[13] Sint, R. et al. *Combining unstructured, fully structured and semi-structured information in semantic wikis*. in *CEUR Workshop Proceedings*. 2009.

[14] Hsueh, P.-Y. et al. *Data quality from crowdsourcing: a study of annotation selection criteria*. in *Proceedings of the NAACL HLT 2009 workshop on active learning for natural language processing*. 2009.

[15] Galvão, J. et al. *Towards Designing Conceptual Data Models for Big Data Warehouses: The Genomics Case*. in *European, Mediterranean, and Middle Eastern Conference on Information Systems*. 2020. Springer.