Luís Carlos Guimarães Lourenço

Licenciado em Ciências da Engenharia Eletrotécnica e de Computadores

Big data analytics for intra-logistics process planning in the automotive sector

Dissertação para obtenção do Grau de Mestre em

Engenharia Eletrotécnica e de Computadores

Orientador:	Ricardo Luís Rosa Jardim Gonçalves, Professor Catedrático, Universidade Nova de Lisboa
Co-orientador:	Ruben Costa, Professor Auxiliar Convidado, Universidade Nova de Lisboa

Júri

Presidente:	Doutor Arnaldo Manuel Guimarães Batista - FCT/UNL
Arguentes:	Doutor André Dionísio Bettencourt da Silva Rocha - FCT/UNL
Vogais:	Doutor Ruben Duarte Dias da Costa - FCT/UNL

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Outubro, 2020

Big data analytics for intra-logistics process planning in the automotive sector

Este documento foi gerado utilizando o processador (pdf)LaTeX, com base no template "novathesis" [1] desenvolvido no Dep. Informática da FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

Lorem ipsum.

Acknowledgements

The conclusion of this project represents the end of a very important chapter of my life and many people deserve to be mentioned.

I would like to thank Professor Ricardo Gonçalves for giving me the opportunity to work on this project.

I would like to thank Professor Ruben Costa for all the guidance provided throughout this dissertation.

The remaining members of the research centre also deserve credit for all the work and insights provided.

A special thanks to Diogo Graça for his mentorship during my internship, and to the entire logistics department in VWAE for all the help provided.

To all my friends and colleagues' thanks for making this journey easier and happier.

Last but not least a very special thanks to my family for all the love and support.

Abstract

The manufacturing sector is facing an important stage with Industry 4.0. This paradigm shift impulses companies to embrace innovative technologies and to pursuit near-zero fault, near real-time reactivity, better traceability, and more predictability, while working to achieve cheaper product customization.

The scenario presented addresses multiple intra-logistic processes of the automotive factory Volkswagen Autoeuropa, where different situations need to be addressed. The main obstacle is the absence of harmonized and integrated data flows between all stages of the intra-logistic process which leads to inefficiencies. The existence of data silos is heavily contributing to this situation, which makes the planning of intra-logistics processes a challenge.

The objective of the work presented here, is to integrate big data and machine learning technologies over data generated by the several manufacturing systems present, and thus support the management and optimisation of warehouse, parts transportation, sequencing and point-of-fit areas. This will support the creation of a digital twin of the intra-logistics processes. Still, the end goal is to employ deep learning techniques to achieve predictive capabilities, all together with simulation, in order to optimize processes planning and equipment efficiency.

The work presented on this thesis, is aligned with the European project BOOST 4.0, with the objective to drive big data technologies in manufacturing domain, focusing on the automotive use-case.

Keywords: Industry 4.0, Data Mining, Machine Learning, Big Data, Digital-Twin

Resumo

O setor de manufatura enfrenta uma etapa importante com a Indústria 4.0. Esta mudança de paradigma impele as empresas a adotar tecnologias inovadoras para atingir falhas quase nulas, reatividade em tempo real, melhor rastreabilidade e previsibilidade, enquanto trabalham para obter uma customização mais barata do produto.

O cenário em estudo aborda vários processos intra-logísticos da fábrica automóvel Volkswagen Autoeuropa, onde diferentes situações necessitam melhoramentos. O principal obstáculo é a ausência de fluxos de dados e integração entre todas as etapas do processo intra-logístico, o que leva a ineficiências. A existência de silos de dados contribui fortemente para estas situações, o que torna o planeamento de processos um desafio.

O objetivo do trabalho apresentado aqui é integrar tecnologias de big data e machine learning nos dados gerados pelos diversos sistemas de produção presentes e, assim, apoiar o gerenciamento e a otimização das áreas de armazém, transporte de peças, sequenciamento e pontos de aplicação. Esta dissertação apoiará também a criação de um gêmeo digital dos processos intra-logísticos, ainda assim, o objetivo final é empregar técnicas de deep learning para obter capacidades preditivas e juntamente com a simulação otimizar o planeamento de processos e a eficiência de equipamentos.

O trabalho apresentado neste documento está embebido no projeto europeu BOOST 4.0, com o objetivo de impulsionar tecnologias de big data no domínio da manufatura, com foco no setor automóvel.

Palavras-chave: Indústria 4.0, Machine Learning, Big Data, Digital Twin

Contents

List of Figures

List of Tables

Acronyms

ACID	Atomicity, Consistency, Isolation, Durability
AGV	Automated Guided Vehicle
ANN	Artificial Neural Network
API	Application Programming Interface
ARMA	Autoregressive Moving Average
BPNN	Back-Propagation Neural Network
CPS	CyberPhysical System
CPU	Central Processing Unit
EFNN	Enhanced Fuzzy Neural Network
ETL	Extract, Transform, Load
IDC	International Data Corporation
IoT	Internet of Things
JSON	JavaScript Object Notation
KPI	Key Performance Indicator
LSTM	Long Short-Term Memory
ML	Machine Learning
MLP	Multilayer Perceptron
MSE	Mean Squared Error
POF	Point Of Fit

RDBMS	Relational Database Management System
RDD	Resilient Distributed Dataset
RFID	Radio Frequency Identification
RNN	Recurrent Neural Network

*

SLA	Service-Level Agreement
SVM	Support Vector Machines
VWAE	Volkswagen Autoeuropa

Chapter 1

Introduction

We live in times of innovation in all fields, and thanks to constant technological developments, globalization, increasing customer expectations and aggressive markets all through the world, companies, business's and academics are working to apply these revolutionary innovations in our advantage.

In the last few years major advances in technologies like internet of things, big data, cloud computing, artificial intelligence and many others are fuelling a new Industrial revolution and in result of that smart manufacturing is becoming the focus of the global manufacturing transformation.

This new revolution is called Industry 4.0 and like in the previous industrial revolutions, technology changed the paradigm, the first one created the mechanization of processes with the steam engine, the second introduced mass production thanks to electricity and the third offered automation due to the introduction of electronic devices. The industrial revolutions and it's driving technologies are represented in figure 1.1 .

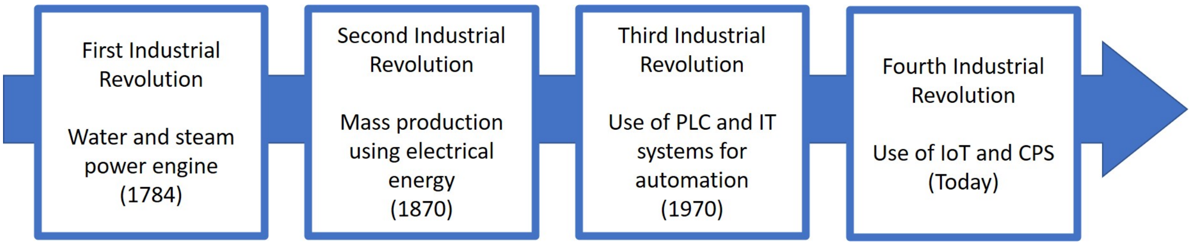

First Industrial Revolution
Water and steam power engine (1784)
Second Industrial Revolution
Mass production using electrical energy (1870)
Third Industrial Revolution
Use of PLC and IT systems for automation (1970)
Fourth Industrial Revolution
Use of IoT and CPS (Today)

Figure 1.1: Industrial revolutions

Today manufacturing industries are changing from the mass production of the past to customized production to meet the growing customer expectations [1].

Another trending topic is Big Data due to the enormous growth of data available in the past few years, a study by IDC (International Data Corporation) titled “data age 2025” predicts that worldwide data creation will grow to an enormous 163 zettabytes by 2025, that is ten times more than 2017.[2]

This paradigm shift brings a huge amount of data associated that needs to be properly processed in order to achieve the desired outcomes. Integrated analysis for the manufacturing big data is beneficial to all aspects of manufacturing.[3] But not just manufacturing can benefit from data all organizations whether large or small, with a data-dependent business model or not can benefit from a better understanding of its data [4]. Big data analytics is a trending subject that companies from all fields are working into their businesses, and with the value and quantity of data rapidly growing every year, we can expect the trend to continue for multiple years making big data an important field to research nowadays.

In the subject of industry 4.0 a technology that also benefits from data growth is Digital-Twin and is likely to become more relevant over the next decade [5]. Digital twin is a live model useful to gather business insights that can be implemented for a specific asset, for an entire facility or even for an individual product.

The concept of machine learning has been around for decades and now it is more relevant than ever because there is an increasing availability of data and computing power, with fast paced developments in the area of algorithms the applications for machine learning in manufacturing will increase [6].

The combination of machine learning and digital-twin can amplify both technologies benefits, where as digital twin can test the accuracy of the machine learning models and different scenarios suggested by the machine learning layers.

The logistics sector does not escape the trends and some major changes are predicted, in fact, logistics represents an appropriate application area for industry 4.0 and its technologies [7]. Logistics has always been a data driven area of business and now more than ever, with the perspectives of real-time tracking of material and product flows[8], improved transport handling, risk management and other features companies need to prepare their logistics departments for the incoming growth of data. “In fact, one could argue that industry 4.0 in its pure vision can only become reality if logistics is capable of providing production systems with the needed input factors . . . ”[7].

This dissertation is embedded in the VWAE (Volkswagen Autoeuropa) pilot of the biggest European initiative in Big Data for Industry 4.0 called BOOST 4.0. This project is funded by the European union and pretends to lead the construction of the industrial European data space and provide the industrial sector with the necessary tools to obtain the maximum benefit from Big Data.

1.1 Problem Description

As previously stated, this dissertation is embedded in the VWAE pilot that deals with improving efficiency on intralogistics operations using new digital technologies, developed under the Industry 4.0 paradigm.
To acquire the necessary knowledge about the business, logistics process and general work of the factory an internship for the duration of the work was seen by both parts as a positive measure and the most fruitful way to proceed.
Intralogistics is the process responsible for components flow since the moment they arrive at the factory to the moment they are applied on the final product. The logistics of delivering parts to the assembly line plays a major role in the success of a car manufacturer and consequently, optimizations to the logistics environment reduce substantially production costs.
Within this project, we started by focusing on a single component, in this case car batteries because this component goes into every car produced, it is reasonably valuable and easy to track. This process is illustrated by figure 1.2.

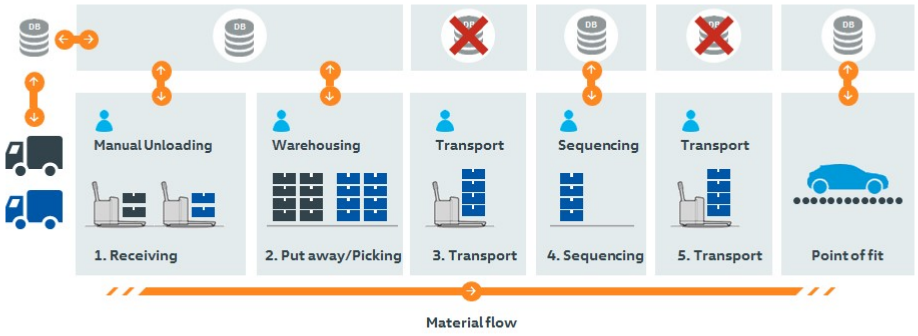


Figure 1.2: current operations of the logistics processes

The processes within intra-logistics on VWAE are the following and they are represented in figures 1.3,1.4 and 1.5 [9]:

- Receiving – On the receiving area trucks are traditionally unloaded by a manually forklift operation, and then the unit loads are transported to the warehouse where they will be stored.

- Warehousing – On the warehouse parts are stored either in shelve or block storage concept. System wise there is one database to control the parts coming from each truck and then a separate database, which registers the unloading, transportation and storing of the material in the warehouse.

- Transport (to sequencing) - An automatic line feeding system based on real vehicle demands generates parts call offs after interacting with real time stock data to replenish the points of use at commissioning areas called SUMA's, or directly at the assembly line, for parts that do not require sequencing, using a pull methodology/-concept. The transport is then made by tow trucks, and the record of these internal transports is stored in a different database. In this process there is an area called Bahnof where parts are placed by the warehouse forklifts to wait for transport to the production line or sequencing.

- Sequencing - The next step will be the picking process for the correct sequencing in the SUMA. Here, the operator follows system electronic picking of parts according to the vehicle sequence on the production line. These operations are executed under the principles of the lean production system[10] [11].
- Transport (to point of fit) – The transport from the sequencing areas to the point of application is made either by AGV's (Automated Guided Vehicles) or again tow trucks. AGV's have data stored in different databases depending on its manufacturer.
- point of fit - Finally, the parts are manually delivery at the point of fit by the line-feeding operator.

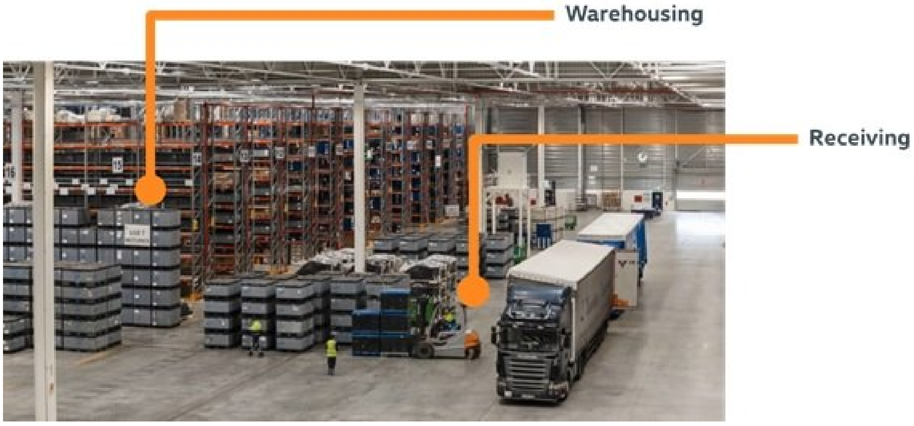


Figure 1.3: Receiving and warehousing at VWAE

Throughout the years, VWAE has done numerous optimizations in its logistics process, namely, with the introduction of AGV's and with the implementation of auxiliary sequencing tools. Having this said, there are still some constraints.
The main issue regarding logistic processes is that there is an absence of a "Big picture", all of the different parts of the process are disconnected on data and on knowledge, there is not an integrated data source nor a single entity with a deep understanding of the whole process.
The lack of communication and integration between the different systems create data silos which makes managing process flows throughout the different steps a challenging task, and the multiple generations of technologies found aggravate this issue since recent systems are prepared for the 4.0 revolution and older ones require multiple steps to even

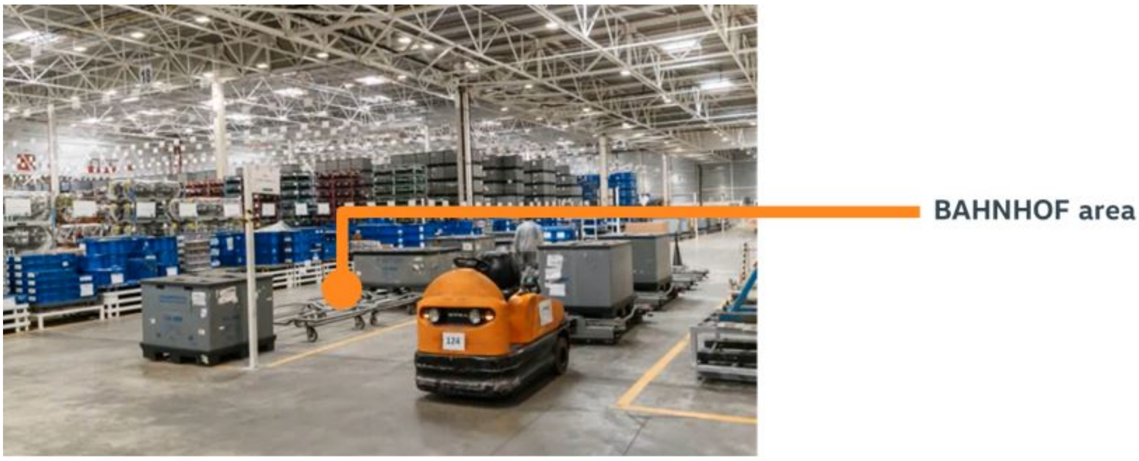
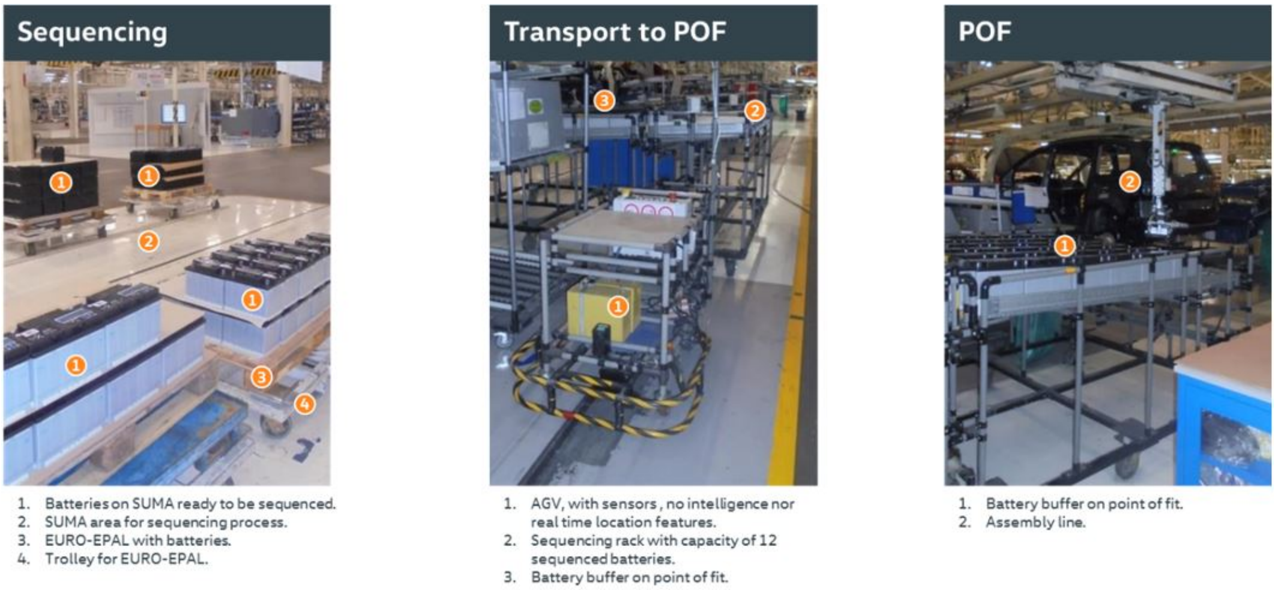


Figure 1.4: Bahnhof area at VWAE



1. Batteries on SUMA ready to be sequenced.
2. SUMA area for sequencing process.
3. EURO-EPAL with batteries.
4. Trolley for EURO-EPAL.

1. AGV, with sensors , no intelligence nor real time location features.
2. Sequencing rack with capacity of 12 sequenced batteries.
3. Battery buffer on point of fit.

1. Battery buffer on point of fit.
2. Assembly line.

Figure 1.5: sequencing and batteries POF at VWAE

gather data. The complexity of the logistics process on a plant of this size and the multi-source, multi-structured, high volume, variety and veracity nature of the data make it very hard to handle, analyse and correlate.

Most of the organizations have huge volumes of structured data housed in different data sources such as mainframes and databases, and also unstructured data sets. Providing integrated data from such a variety of sources is prerequisite for effective business intelligence[12].

Gathering data from heterogenous sources and manipulating them to prepare for big data analysis processes is still a big problem[13]. The logistics department at VWAE suffers from the absence of any predictive and adaptive functionalities and that force logistics planners to use conventional methods, relying on past experience and trial and error for every decision they make. This reduces the ability of optimizing the system because it takes considerable time, effort and, until deployment, there is no way of validating these changes or to predict their outcomes with an acceptable degree of confidence.

Data errors are present with some regularity, due to lack of both data validation and awareness of the importance of data validity. This problem reduces the confidence of both the decision makers and planners at VWAE in the data available which leads to its lack of use and value.

One important aspect of logistics in manufacturing is warehouse management. Managing a warehouse is very complex because of the multiple variables to consider like physical space, internal transport times, inventory costs, security and material quality to name a few. The warehousing management is also constantly under pressure to feed the production line because stoppages can be very expensive.

Warehouse management at VWAE is no different and must account for all these variables for a few thousand different parts with very different characteristics and processes associated. For our selected part, car batteries, inventory management is especially important for multiple reasons, it's a valuable component so stall money is a factor, it has to be stored in ground level and does not support stacking of other packages above so it occupies premium warehouse location and they also have expiration dates. There is a necessity to be more efficient in warehousing space, stall money, transport costs and pollutant emissions and that lead to a necessity of optimizing inventory levels.

1.2 Research question and hypothesis

Regarding the problem mentioned above, a question can be asked: "How to optimize the inventory levels based on the production?" The hypothesis to prove is that the inventory levels can be optimized with a data driven system that analyses the available data from stock, supply and demand, and learns from the data to provide optimizations.

1.3 Proposed solution

The proposed solution consists of a data-driven system that can collect, clean, prepare and integrate data from multiple sources, learn from historical data and suggest optimizations.

As a primary objective we intend to minimize situations of overstock by reducing order quantities and number of trucks.

To that end I will utilize historical data from stock levels to analyse the magnitude of the problem and where are opportunities for improvements. Then I will use data from the production line to predict the usage of each material to calculate the optimal number of inventory is at a given moment.

We intended to implement a multi-layer architecture, which for this use-case wont necessary be a "big data" architecture but will be built with a future integration in a cloud computing system like apache spark in mind for the purpose of scalability. The multiple layers will be described in detail throughout this dissertation. The said layers are the following:

1. ETL (extract, transform, load) – To implement a data driven system its necessary to extract the data from its source, transform it into a clean structured format to load into said system. This layer consists of multiple operations of loading, cleaning, reshaping, resampling and normalization to get the data from its source into our machine learning (ML) layer or directly to the processing or presentation layer.

2. Storage layer – this layer consists of a database with the clean, structured and integrated data necessary to save historical data.

3. ML layer – This layer consists of a machine learning model that receives the already prepared data and learns patterns, behaviours, and trends without being explicitly programmed from that data to forecast the future states of that said data. Those forecasted values are then forward to the next layer.

4. Processing – This layer consults the available data, including the forecasts provided by the ML layer, and calculates the suggested values of each target.

5. Presentation – This layer consists of ways to present data and insights to the planners, by the creation of graphs and data tables in intuitive ways. With this addition decision makers are equipped with data and insights to make the best possible decision.

This architecture was chosen because the main problem encountered was the absence of a "Big Picture" of the data available, because of the existence of data silos and lack of data validation. This way our ETL layer can eliminate the data silos by integrating the data and cleaning and structuring the data in the process. The machine learning layer can be used to learn from data and output insights, in this use case it will learn from historical data to forecast the next 5 days of consumption of car batteries by the production line. The processing is necessary to interpret the output of the machine learning layer, and finally the presentation layer intends to solve the absence of the "Big Picture" by presenting the data from all clusters in the same platform.

1.4 Methodology

This dissertation will follow a CRISP-DM (Cross-industry standard process for data mining) reference model. Within this model the life cycle of a data mining project is broken down in six phases which are shown in figure 1.6. [14]

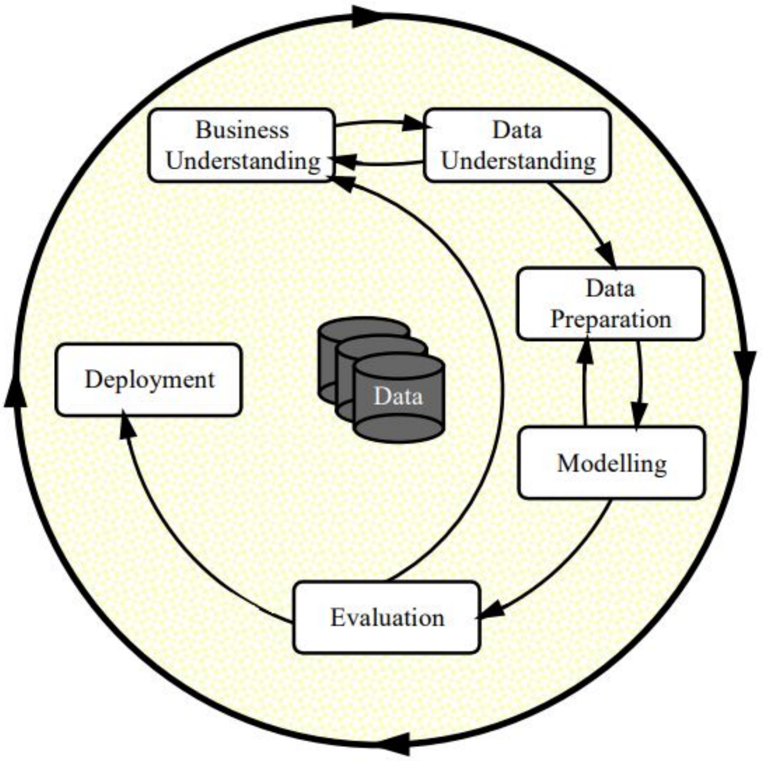


Figure 1.6: Phases of CRISP-DM model

1. Business understanding – Vital to know what to look for and defining objectives

2. Data understanding – Considerations on how to integrate data from multiple sources.

3. Data preparation – Covers all activities to construct the final dataset

4. Modelling – Decide and apply the modelling techniques.

5. Evaluating – Evaluate the results obtained in the previous step and decide new objectives and future tasks.

6. Deployment – Define a strategy to implement the results validated on step 5.

This methodology does not have a strict sequence of the phases and moving back and forth between phases is always required in order to improve the outcome iteratively.
To acquire the necessary knowledge about the business, logistics process and general work of the factory an internship for the duration of the work was seen by both parts as a positive measure and the most fruitful way to proceed.

1.5 BOOST 4.0 Contributions

As previously stated, this work is integrated in the VWAE pilot of the BOOST 4.0 European project. This section will provide a description of the pilot and its objectives as well as description of the different phases of the pilot and the contributions of this thesis to the project.

Boost 4.0 is seeking to improve the competitiveness of Industry 4.0 and to guide the European manufacturing industry in the introduction of Big Data in the factory, along with the necessary tools to obtain the maximum benefit of Big Data. In respect to global standards, Boost 4.0 is committed to the international standardization of European Industrial Data Space data models and open interfaces aligned with the European Reference Architectural Model Industry 4.0 (RAMI 4.0).

The standardization of industry 4.0 compliant systems or smart manufacturing systems include many aspects [15]. Future smart manufacturing infrastructures must enable the exploitation of new opportunities. Even today, people are surrounded by interconnected digital environments continuously generating more synergies with connected devices and software. Such an evolution happens also in the manufacturing domain as in Volkswagen. Future Smart Manufacturing infrastructures are confronted with the digitalisation and virtualisation of (physical) objects enhanced with sensors, processors, memory and communication devices, able to communicate coactively and to exchange information independently through a reactive, predictive, social, self-aware and/or autonomous behaviour [16] [17]. A used term for such intelligent physical objects is Cyber-Physical System (CPS) which are communicating in (Industrial) Internet of Things ((I)IoT) networks.

To exploit new opportunities, specific requirements as real-time, security or safety have to be considered. Smart Manufacturing infrastructures have to be based on network technologies which enable a secure (encryption, authentication, robustness, safety), vertical and horizontal cross-domain and cross-layer communication between stationary and mobile objects (as virtual objects, sensors, actors, devices, things or systems). Network technologies must comply with specific requirements related to e.g. real-time, safety, security, data amounts, wired or wireless, passive, or active, etc.[18]. Lower level fields (process control or real-time statistics) require time frame abilities of seconds or even milliseconds for response, whereas higher levels (production planning or accounting) only require time frames of weeks or months[18]. Architectures, as the RAMI 4.0 in figure 1.7, in general, are describing the ordering of components/modules and their interaction and should provide a unified structure and wording for used terms. An architecture should include a logical, a development, a process and a validation view, and should provide scenarios for a validation as proposed by Philippe Kruchten in his 4+1 architectural view model [19]. A smart manufacturing architecture should also provide a unified structure and wording covering mandatory aspects in smart manufacturing as product, system or order life cycles, value streams, information flows, or hierarchical layers. Such architectures are currently under development. (Physical) reachable objects inside a smart manufacturing network (e.g. digitalised and virtualised field level devices, systems, material, integrated humans, virtual concepts (e.g. of products in the design phase), etc.), have to fulfil a range of requirements. Objects should communicate using a unified communication protocol, at least at the application level, and should be based on a unified semantic to enable a mutual identifiability and understanding. The object itself should provide its own features as a service (e.g. state information or functionalities) and should be able to provide its own description next to extended information as manuals, specifications or wear information. All these have to be kept next to further requirements related to security, safety or quality of service [20] [21]. Finally, various applications that use services of deployed objects to realise e.g. control systems, systems of systems through service orchestration, or - as focused in in this work - Big Data analysis applications can be implemented. Standards can be classified according to what role they play in the system architecture. At this stage in the Boost 4.0 project we have the RAMI 4.0 (fig 1.7). In figure 1.8 we have the Boost 4.0 architecture. Here we can see Boost 4.0 horizontal layers, visualization, data analytics, data processing, data management and external Data sources/Infrastructure as well as the Vertical layers, development, communications and connectivity, data sharing platforms and privacy/security. This dissertation will contribute mainly on the visualization, data analytics, data processing, data management and external data sources/Infrastructure layers.

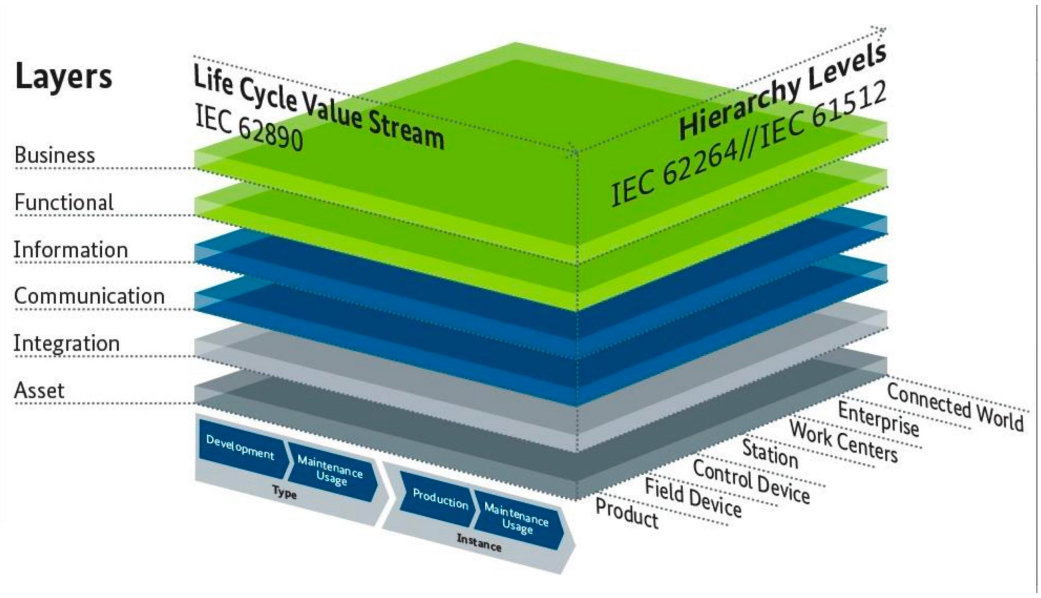


Figure 1.7: RAMI 4.0 Reference Architecture

This pilot is structured in four different phases as we can see on figure 1.9.

Phase 1 - The initial version of the pilot's implementation mainly comprised the overall test of the closeness of the simulation with the reality of the logistics operations at VWAE.

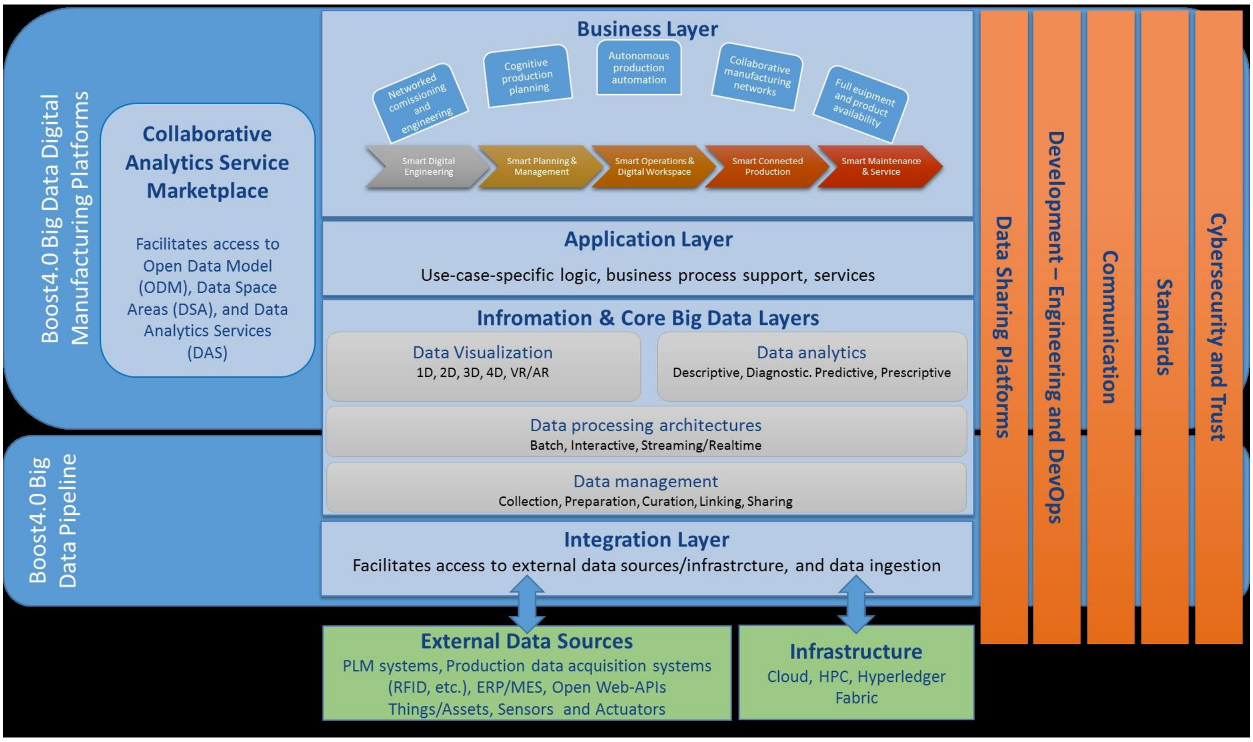

Boost4.0 Big Data Digital Manufacturing Platforms
Collaborative Analytics Service Marketplace
Facilitates access to Open Data Model (ODM), Data Space Areas (DSA), and Data Analytics Services (DAS)
Boost4.0 Big Data Pipeline
Business Layer
Networked comissioning and engineering
Cognitive production planning
Autonomous production automation
Collaborative manufacturing networks
Full euipment and product availability
Smart Digital Engineering
Smart Planning & Management
Smart Operations & Digital Workspace
Smart Connected Production
Smart Maintenance & Service
Application Layer
Use-case-specific logic, business process support, services
Infromation & Core Big Data Layers
Data Visualization
1D, 2D, 3D, 4D, VR/AR
Data analytics
Descriptive, Diagnostic. Predictive, Prescriptive
Data processing architectures
Batch, Interactive, Streaming/Realtime
Data management
Collection, Preparation, Curation, Linking, Sharing
Integration Layer
Facilitates access to external data sources/infrastrcture, and data ingestion
External Data Sources
PLM systems, Production data acquisition systems (RFID, etc.), ERP/MES, Open Web-APIs Things/Assets, Sensors and Actuators
Infrastructure
Cloud, HPC, Hyperledger Fabric
Data Sharing Platforms
Development – Engineering and DevOps
Communication
Standards
Cybersecurity and Trust

Figure 1.8: Boost 4.0 Architecture

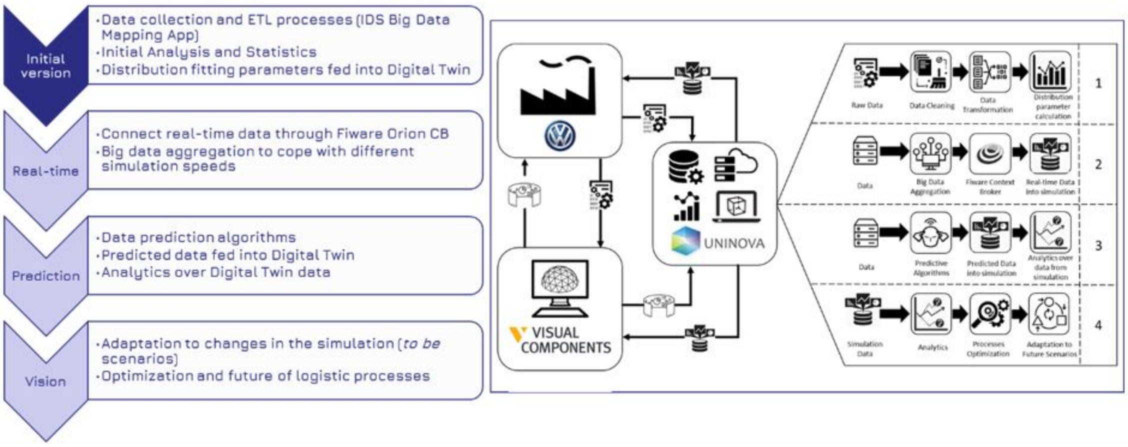

Initial version
·Data collection and ETL processes (IDS Big Data Mapping App)
·Initial Analysis and Statistics
·Distribution fitting parameters fed into Digital Twin
Real-time
·Connect real-time data through Fiware Orion CB
·Big data aggregation to cope with different simulation speeds
Prediction
·Data prediction algorithms
·Predicted data fed into Digital Twin
·Analytics over Digital Twin data
Vision
·Adaptation to changes in the simulation (to be scenarios)
·Optimization and future of logistic processes
UNINOVA
VISUAL COMPONENTS
Raw Data
Data Cleaning
Data Transformation
Distribution parameter calculation
1
Data
Big Data Aggregation
Fiware Context Broker
Real-time Data into simulation
2
Data
Predictive Algorithms
Predicted Data into simulation
Analytics over data from simulation
3
Simulation Data
Analytics
Processes Optimization
Adaptation to Future Scenarios
4

Figure 1.9: Pilot structure and phases

This phase was divided into several iterations in order to have the best possible fitting between what is being simulated in Visual Components "Digital-Twin" and the reality in terms of logistics processes, in accordance to the main business scenarios for this pilot. The tasks of this phase are represented in figure 1.10 and this dissertation contributed to the data cleaning and data transformation tasks.

Phase 2 - Real-Time Scenario. In this phase, real-time data is fed into the simulation in order to confirm that the simulation clearly depicts the real-world processes, to validate if the real-world processes can be optimized and also to check the as-is situation

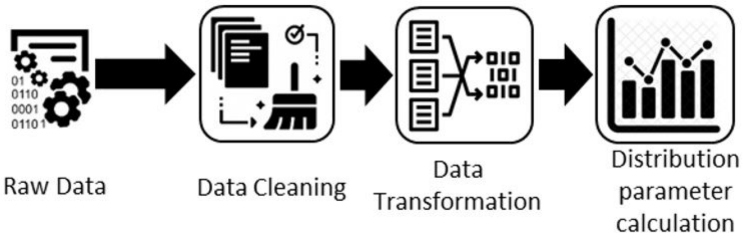


Figure 1.10: phase 1

when tweaks in the actual process are performed in the simulation. The data is fed to the simulation environment through a Publish-Subscribe mechanism, as is the case of the OPC-UA standard or the FIWARE ORION Context Broker, meaning that when a set of data is published into the service, the simulation environment will get it through its subscription to the Pub-sub service. On figure 1.11 we have the tasks of this phase, this dissertation contributed to the Big data aggregation.

Phase 3 Prediction. This phase is characterized by the use of Data Mining and Machine

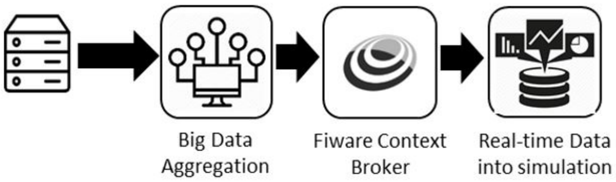


Figure 1.11: phase 2

Learning algorithms, both for prediction of future data values and on the analysis of data retrieved from the simulation environment. The first process will be to predict future data depending on specific tweaks to the processes, whether they are made directly on the physical dimension of the simulation (e.g. placing the sequencing area in a different place, changing an human operator for an AGV or robot, etc.) or on the data per se (e.g. increase the number of jobs in the Point-of-fit, increase the time intervals between truck arrivals, etc.). The predicted data would then be fed into the simulation environment, in order to check the impact of the tweaks in the logistics operation, i.e. if the whole process would still correspond to the necessary production requirements or not. In the case the process does not meet the necessary requirements, then solutions for the encountered issues must be found. The tasks of this phase are in figure 1.12 and this dissertation contributed to the predictive algorithms task.

Phase 4 Future Digital Twin. This phase comprises the final version of the digital twin, in which the future operations of the logistics area at VWAE will be tested prior to real implementations. This process of digital twin testing will provide a solid ground to create new processes, optimize existing ones and test significant changes in the overall logistics operation without the need of performing real-world pilots, saving money, time and human resources that would otherwise be needed to perform such piloting activities. This is

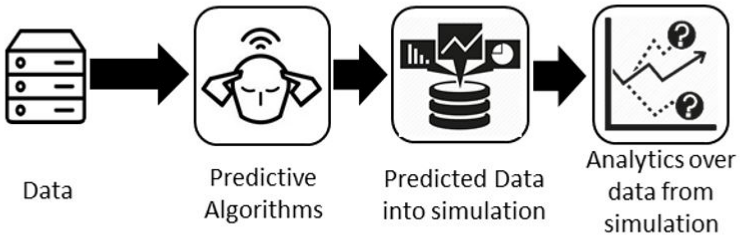

Data
Predictive Algorithms
Predicted Data into simulation
Analytics over data from simulation

Figure 1.12: phase 3

crucial to VWAE since, up to now, the only way to perform testing activities is to couple them into the everyday operation, which brings serious problems in terms of execution times, resource usage and return of investment. Furthermore, when piloting some optimization of processes or the assessment of the use of new technologies on the logistics operations does not meet the required expectations, all of the above problems are even more critical, since the effort spent in the tests, regarding money, time and resources does not contribute to a substantial improvement of the operation. The tasks of this phase are in figure 1.13 and this dissertation contributed to the analytics task.

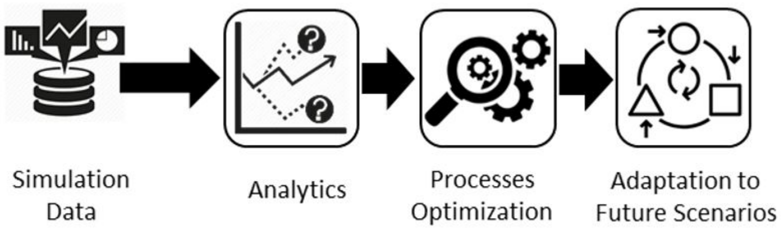


Figure 1.13: phase 4

1.6 Thesis outline

This dissertation is divided in 6 chapters. The first chapter is the introduction one, it frames and describes the problem at study and presents a conceptualization and description of the proposed solution. Literature review and study of previous solutions and research on similar problems and key technologies forms the next chapter. After this a chapter with the description of the data available and used throughout the dissertation. The architecture of the system is the next chapter and it is divided into sub-chapters for each step of the process, followed by a chapter presenting the results obtained. Finally, a chapter with conclusions drawn from the work and a description of the work that can derive from the one present here.

Chapter 2

State of the Art

This chapter contains a review of the concepts and technologies addressed in this dissertation.

2.1 Machine Learning

Machine learning (ML) is the study of algorithms and statistical models that computers use to perform specific tasks without using explicit instructions, relying on patterns and inference instead. It focuses on the development of computer programs that access data and use it to learn (training data). This technology is seen as a subset of artificial intelligence.

There are multiple categories of machine learning and each one of these differ in approach, type of data input and output, and the type of problem that they are intended to solve. We can separate machine learning in 3 main categories:

- Supervised Learning – Builds a mathematical model of a set of data that contains both the inputs and the desired outputs. To every output there is one or multiple inputs. Through iterative optimization of an objective function, the system can provide outputs to any new inputs after enough training. Supervised learning algorithms include classification, for cases where the outputs are restricted to a limited set of values, and regression for cases when the outputs can have any numerical value.

- Unsupervised learning – Is used when the data provided is neither classified nor labelled, and instead of figuring out the right output identifies common features within the data and can infer functions to describe hidden structure on data.

- Reinforcement learning – Ought to take actions in an environment to maximize a notion of cumulative reward. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. It allows software agents to autonomously determine the ideal behaviour in each scenario. These algorithm's do not assume the knowledge of an exact mathematical model and are used when exact models are infeasible.

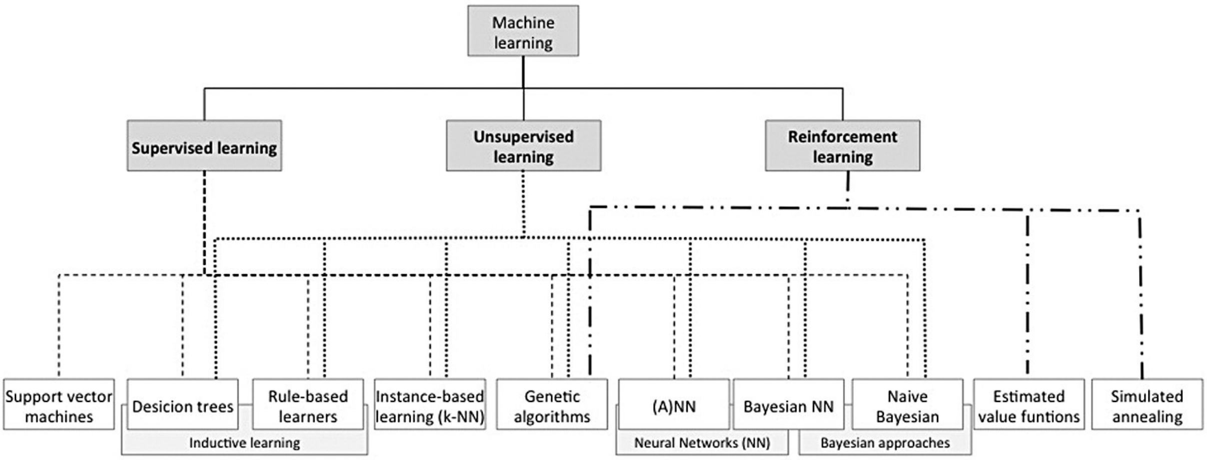


Figure 2.1: Structuring of machine learning techniques and algorithms [6]

Figure 2.1 shows the structuring of machine learning techniques and algorithms, and that in all categories there is a big range of different algorithms, each one with its advantages and disadvantages.

Machine learning has applications in multiple areas, approaches for predicting future inbound logistic processes already exist[22], forecasting of supply chains showed improvements and increased adaptability with the use of machine learning algorithms[23], in healthcare machine learning algorithms proved successful in predicting early colorectal cancer metastasis using digital slide images[24].

In industry, supervised machine learning techniques are mostly applied due to the data-rich but knowledge-sparse nature of the problems [25]. The general process contains several steps handling the data and setting up the training and test dataset by the teacher, hence supervised [26].

In 2017 an article[27] implemented a machine learning based system to respond to a problem of optimal order placements in electronic equity markets and achieved substantial reductions of transactions costs.

Multiple machine layer techniques are being applied with success on scheduling problems like this article [28] that proposes a framework to optimize scheduling of processes in order to reduce power consumption in data-centre's, they utilize machine learning techniques to deal with uncertain information and use models learned from previous system behaviours in order to predict power consumption levels, CPU (Central Processing Unit) loads, and SLA(service-level agreement) timings, and improve scheduling decisions.

Machine learning algorithms are becoming more and more useful with the growth of big data, since it is not possible or practical to have programmer's constantly adapting code to extract useful information from data. There are multiple examples of cases where the use of big data techniques aided by machine learning produced valuable results in varied areas like energy, logistics, agriculture, marketing or even health. A good example are search engines that use ML algorithms to recommend advertisements related with the content searched [29].

2.1.1 Predictive techniques

There is a wide range of predictive techniques and mainly two categories, regression techniques and machine learning ones. With regression models the focus lies on establishing a mathematical equation as a model to represent the interactions between the different variables in consideration. Depending on the situation there are a wide variety of models that can be applied while performing predictive analytics.

One of this models is the linear regression model that analyses the relationship between the response or dependent variable and a set of independent or predictor variables. This relationship is expressed as an equation that predicts the response variable as a linear function of the parameters. These parameters are adjusted so that a measure of fit is optimized. Much of the effort in model fitting is focused on minimizing the size of the residual, as well as ensuring that it is randomly distributed with respect to the model predictions. A proposed local linear regression model was applied to short-term traffic prediction in this paper[30] and the performance of the model was compared with previous results of nonparametric approaches that are based on local constant regression, such as the k-nearest neighbour and kernel methods, by using 32-day traffic-speed data collected on US-290, in Houston, Texas, at 5-min intervals. It was found that the local linear methods consistently showed better performance than the k-nearest neighbour and kernel smoothing methods.

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. Although it's a simple model in some cases it can outperform more advanced models. This study [31] uses Logistic Regression, Moving Average and BPNN (Back-Propagation Neural Network) methods for sales models designed to predict daily fresh food sales and found that the correct percentage obtained by the logistic regression to be better than that obtained by the BPNN and moving average models

Machine learning was already described before and some of its techniques can be used to conduct predictive analytics.

Neural networks are nonlinear sophisticated modelling techniques that are able to model complex functions. They can be applied to problems of prediction, classification or control in a wide spectrum of fields and are used when the exact nature of the relationship between inputs and output is not known. A key feature of neural networks is that they

learn the relationship between inputs and output through training.
The multilayer perceptron (MLP) consists of an input and an output layer with one or more hidden layers of non-linearly-activating nodes or sigmoid nodes. This is determined by the weight vector and it is necessary to adjust the weights of the network. The backpropagation employs gradient fall to minimize the squared error between the network output values and desired values for those outputs. The weights adjusted by an iterative process of repetitive present of attributes. Small changes in the weight to get the desired values are done by the process called training the net and is done by the training set (learning rule).
Support vector machines (SVM) are used to detect and exploit complex patterns in data by clustering, classifying and ranking the data. They are learning machines that are used to perform binary classifications and regression estimations. They commonly use kernel-based methods to apply linear classification techniques to non-linear classification problems. There are several types of SVM such as linear, polynomial, sigmoid etc. Multiple authors written about this. This paper [32] proposes an support vector machine model to forecast the streamflow values of Swan River near Bigfork and St. Regis River near Clark Fork of Montana, United States and this model outperformed the other models tested, the autoregressive moving average model (ARMA) and an artificial neural network (ANN). Another SVM model was successfully utilized to predict a daily electricity price forecast on this paper[33].

2.1.2 LSTM

Long short-term memory (LSTM) networks are a type of RNN and were discovered in 1997 by Hochreiter and Schmidhuber and set accuracy records in multiple applications domains. [34]
LSTM are deep learning systems that avoid the vanishing gradient problem which means that prevent backpropagated errors from disappearing or overgrowing . LSTM are normally augmented by recurrent gates called “forget gates”. [35]. So, errors can flow backwards through unlimited numbers of virtual layers unfolded in space. LSTM can learn tasks that require memories of events that happened thousands or even millions of discrete time steps earlier. [36] LSTM differ from other networks because they can work with long delays between events and mainly because they can handle high and low frequency events at the same time.
Multiple authors are using LSTM to make predictions to important datasets, a paper [37] proposed an approach to forecast PM2.5 (Particulate Matter) concentration using LSTM by exploiting Keras[38], which is a high-level neural networks API written in Python and capable of running on top of Tensorflow, to build a neural network and run RNN with LSTM through Tensorflow. The results showed that the proposed approach can effectively forecast the value of PM2.5.

Another paper [39] modelled and predicted China stock returns using LSTM. The historical data of China stock market were transformed into 30-days-long sequences with 10 learning features. That LSTM model compared with random prediction method improved the accuracy of stock returns prediction.

LSTM models accept multiple input and output types of data, one example of that is a paper [40] that introduced an algorithm of text-based LSTM networks for automatic composition and reported results for generating chord progressions and rock drum tracks. The experiments show LSTM provides a way to learn the sequence of musical events even when the data is given as text and the authors plan to examine a more complex network with the capability of learning interactions within music (instruments, melody/lyrics) for a more complete automatic composition algorithm.

2.2 Big Data Analytics

With the exponential growth in the volume of data produced, big data is a concept whose relevance has grown, a tendency with no signs of slowing down in a near future. In general, big data is used to describe a large amount of structured, semi-structured and unstructured data created by data sources, which would need too much time and money to be stored and analysed. Big data can also be defined by the four characteristics, also named “the four V's”: [3]

- Volume, for the scale of the data produced, which makes it difficult to be processed by regular data processing techniques.
- Velocity, by the pace at which the data is produced, demanding a much higher processing capacity.
- Variety, in terms of content, format and size, which does not enable a standard method for processing all the data.
- Value of the hidden information that can be collect by analysing such a large amount of data.

Data Analytics may correspond to the application of tools and techniques to extract insights and knowledge from data, by analysing it through any of Statistics, Data Mining and Machine Learning techniques. Although statistical analytics is supported by well-known statistical techniques, which are more easily deployed on a Big Data context, in the case of Data Mining and Machine Learning, the passage to a Big Data environment is not a trivial task, since it comprises the reconfiguration of algorithms to be deployed in Big Data execution engines.

In typical data mining systems, the mining procedures require computational intensive computing units for data analysis and comparisons. A computing platform is, therefore, needed to have efficient access to, at least, two types of resources: data and computing processors.

For Big Data mining, because data scale is far beyond the capacity that a single personal computer can handle, a typical Big Data processing framework will rely on cluster computers with a high-performance computing platform. The role of the software component is to make sure that a single data mining task, such as finding the best match of a query from a database with billions of records, is split into many small tasks each of which is running on one or multiple computing nodes [41].

Big Data Analytics refers to the implementation of analytic tools and technologies within the scope of Big Data [9]. Hence, Big Data Analytics may be described by two specific concepts, Big Data + Analytics, and the interactions between technologies supporting both concepts.

So, why merge these concepts [42]? First, Big Data provides gigantic statistical samples, which enhance analytic tool results. In fact, the general rule is that the larger the data sample, the more accurate are the statistics and other products of the analysis. Second, analytic tools and databases can now handle big data, and can also execute big queries and parse tables in record time. Moreover, due to a precipitous drop in the cost of data storage and processing bandwidth, the economics of analytics is now more embraceable than ever.

The manufacturing sector is also implementing Big Data Analytics, this paper [43] proposes a big data driven analytical framework to reduce the energy consumption and emission for energy-intensive manufacturing industries. Then an application scenario of ball mills in a pulp workshop of a partner company is presented to demonstrate the proposed framework. The results show that the energy consumption and energy costs are reduced by 3% and 4% respectively.

According to [44] the semiconductor manufacturing industry has been taking advantage of the big data and analytics evolution by improving existing capabilities such as fault detection, and supporting new capabilities such as predictive maintenance. For most of these capabilities, data quality is the most important big data factor in delivering high quality solutions and incorporating subject matter expertise in analytics is often required for realizing effective on-line manufacturing solutions. In the future, an improved big data environment incorporating smart manufacturing concepts such as digital twin will further enable analytics; however, it is anticipated that the need for incorporating subject matter expertise in solution design will remain.

Internet of Things generated data is characterized by its continuous generation, large amount, and unstructured format. The existing relational database technologies are inadequate to handle such IoT generated data because of the limited processing speed and the significant storage-expansion cost, to counter that a paper [45] proposes a sensor-integrated radio frequency identification (RFID) data repository-implementation model using MongoDB and show that the proposed design strategy, which is based on horizontal data partitioning and a compound shard key, is effective and efficient for the IoT generated RFID/sensor big data.

In this paper[46], an overall architecture of big data-based analytics for product lifecycle (BDA-PL) was proposed. It integrated big data analytics and service-driven patterns that helped to overcome the lack of complete data and valuable knowledge. Under the architecture, the availability and accessibility of data and knowledge related to the product were achieved. Focusing on manufacturing and maintenance process of the product lifecycle, and the key technologies were developed to implement the big data analytics. The presented architecture was demonstrated by an application scenario, and the results showed that the proposed architecture benefited customers, manufacturers, environment and even all stages of product lifecycle management, and effectively promoted the implementation of cleaner production.

Big Data in supply chain problems makes it possible to analyse the data at a more advanced level than traditional tools, allowing the processing and combining of data collected from several systems and databases in order to provide a clear picture of the situation. It can provide information on potential interference with the supply chain through the collection and evaluation of data, it is possible not only to protect but also improve the efficiency of the supply chain. This way, interruptions on production are avoided and operational efficiency is increased. Big Data enables the optimization of logistic processes while making the supply chain less prone to failures [47].

2.2.1 Existing Big Data Technologies

There are several surveys, starting from early 2000's up to today, regarding Big Data Analytics. These surveys often describe the same Big Data technologies, which have been evolving throughout the years, coupled with Analytics techniques. The following paragraphs present the most prevalent technologies and tools on all the surveys[42] [48] [49] [50].

Regarding execution engines, the following are the most referred to in literature. Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures. It builds over a data processing paradigm called MapReduce. The MapReduce workflow looks like this: read data from the cluster, perform an operation, write results to the cluster, read updated data from the cluster, perform next operation, write next results to the cluster, etc., Apache Spark is a general-purpose cluster computing engine which is very fast and reliable [51] that started as a research project at the UC Berkeley AMPLab in 2009, and was open sourced in early 2010. Many of the ideas behind the system were presented in various research papers over the years.

Spark offers an abstraction called resilient distributed datasets (RDDs) to support these

applications efficiently. RDDs can be stored in memory between queries without requiring replication. Instead, they rebuild lost data on failure using lineage: each RDD remembers how it was built from other datasets (by transformations like map, join or groupBy) to rebuild itself. RDDs allow Spark to outperform existing models by up to 100x in multi-pass analytics. Spark showed that RDDs can support a wide variety of iterative algorithms, as well as interactive data mining and a highly efficient SQL engine called Spark SQL, which enables queries in SQL to be executed on NoSQL environments. While MapReduce operates in steps, Spark operates on the whole data set in one fell swoop. Spark completes the full data analytics operations in-memory and in near real-time and Spark also works for both batch offline data processing and online stream processing, through its real-time counterpart: Spark Streaming.
Apache Spark also has a Machine Learning library called MLlib [52], which include:

- Classification: logistic regression, naive Bayes;
- Regression: generalized linear regression, survival regression;
- Decision trees, random forests, and gradient-boosted trees;
- Recommendation: alternating least squares (ALS);
- Clustering: K-means, Gaussian mixtures (GMMs);
- Topic modelling: latent Dirichlet allocation (LDA);
- Frequent item sets, association rules, and sequential pattern mining;

Beyond Big Data execution engines, storage and query systems for Big Data also had an enormous evolution in the past few years. MongoDb[53], Apache Cassandra[54] two different storage engines which do not rely on traditional RDBMS (Relational Database Management System) and SQL technologies. Instead, each use a specific type of data storage mechanism. Mongo is based on a document structure, relying on JSON(JavaScript Object Notation) formatted documents to store data. Cassandra is also supported by a file storage system, while HBase maintains the traditional tabular form, used in RDBMS systems. Because most companies are used to using SQL query tools in order to perform complex queries on their systems, several abstractions to NoSQL technologies were added, in order to provide SQL query functionality to these systems.

2.3 Inventory Management

This section describes relevant literature to frame the phenomenon's in focus on this dissertation.
Inventory management is a challenging problem in supply chain management and inventory is the supply of raw materials that an organization maintains to meet its operational

needs. Inventory is defined as a stock of goods that is maintained by a business in anticipation of some future demand. The quantity to which inventory must fall in order to signal that an order must be placed to replenish an item[55].

Capacity and inventory management are key to operations management, as they concern the planning and control of the supply or processing side of matching supply and demand, and because of that they are very a researched area[56].

As this is an old problem multiple authors have written about it.

Dolgui and Prodhon [57] have focused on the development of MRP software for an uncertain environment and have shown that various techniques such as safety stock, safety lead time, and lot-sizing rules can be used to control the supply variability in order to lead the better anticipation of uncertainties.

Multiple authors have showed that safety lead time can be used to work around supply uncertainties like late deliveries[58]. Axsäter in 2006 compared two types of lead times. The comparison showed that inventory levels vary less in case of independent lead times than dependent ones[59]. There is tremendous need for scalable supply chain optimization algorithms to respond to dynamic information, that is, to perform data-driven re-optimization in a timely manner[56]. The classical systems work but they usually assumed stationary demand distributions, but when the demand environment is non static or unknown, optimal policy is often difficult to identify and even when one can identify some solutions those are likely to require a lot of computational power. To tackle these complex optimization problems on an industrial scale, machine learning techniques can be used to generate quick heuristics. Machine learning can provide general purpose algorithms that can readily be applied to multiple different problems without years of specialized research to tailor the solution approach although the generic nature of these algorithms can have worse performances than the ones specifically designed for the problem at hand[60].

According to this paper [61] determining the adequate stock levels balances the overstocking costs, these include costs for holding the safety stocks, for occupying additional storage space and transportation and the Costs of lost sales or production. To deal with this costs the use of data mining techniques ensures that each inventory point (internal warehouse, work-in-process, distribution center, retail store) has the optimal stock levels. Commonly, managers have relied on a combination of ERP (Enterprise Resource Planning), supply chain, and other specialized software packages, as well as their intuition to forecast inventory. However, in today's high uncertain environment and large quantities of disparate data demands new approaches for forecasting inventory across the entire chain. Data mining tools can be used to accurately forecast products to where they are needed.

This paper [62] proposes a hybrid deep learning models for inventory forecasting. According to the highly nonlinear and non-stationary characteristics of inventory data, the models employ Long Short-Term Memory (LSTM) to capture long temporal dependencies and Convolutional Neural Network (CNN) to learn the local trend features and although building CNN-LSTM network architecture and tuning can be challenging, experimental results indicate that the evolved CNN-LSTM models are capable of dealing with complex nonlinear inventory forecasting problem.

This paper [63] from 2019, proposed an inventory forecasting solution based on time series data mining techniques applied to transactional data of medical consumptions because one of the factors that often result in an unforeseen shortage or expiry of medication is the absence of, or continued use of ineffective, inventory forecasting mechanisms. Unforeseen shortage of perhaps lifesaving medication potentially translates to a loss of lives, while overstocking can affect both medical budgeting as well as healthcare provision. The results from this work evidently suggest that the use of data mining techniques could prove a feasible solution to a prevalent challenge in medical inventory forecasting process.

Some authors argue that human decision making, augmented by data-driven decision models suggesting real-time actions, will remain the desired approach for complex operations as algorithms rarely can anticipate all possibilities economically[56]. So, it is desirable that the human be the ultimate decision maker as experience, common sense, intuition, and insights derived from structured models can rarely be replaced by a fully automated solution. Intervention is desired, if not necessary, when we detect stupid solutions due to input or algorithm errors. Analysing their output can generate insight into which variables are significant and which can be ignored and can help enriching analytical models to develop deeper insights, as illustrated by Gijsbrechts [60].

A paper in 2008 [64] develop an enhanced fuzzy neural network (EFNN) based decision support system for managing automobile spares inventory in a central warehouse. And in that system, the EFNN is utilized for forecasting the demand for spare parts. That system when evaluated with real world data outperformed five other models.

Chapter 3

Intralogistics Data Analysis

This chapter will include a description and analysis of the whole range of intralogistics data generated at the VWAE automotive factory.

The objective of this thesis and the BOOST 4.0 project is to contribute to the optimizations of intralogistics processes by applying emerging technologies and take advantage of the data available. To ensure this a data assessment to shed light on how to integrate data in order to achieve a "big picture" of the intralogistics process.

One of the first tasks I set out to accomplish was an overview of the data available. A description of the data available by each cluster.

Note that some of the data is not described in detail to prevent any issues of data protection and confidentiality. For the same reasons some data is described but the sample presented contains less data than the described.

3.1 External Transports Data

The external material transports in this context means transports that start outside the factory. Most of the incoming parts arrive by truck, the available data consisted of raw excel files with tabular data from each truck arriving at the factory.

These files contained timestamps of multiple events for each truck, like arriving time, start and end of unloading, information about the material unloaded like quantity and description, licence plate of the truck and the unloading position as well as transport identification number and material order number. The licence plate, transport identification number and material order number fields are important to the data integration process because these fields represent the same information on the receiving cluster, thus enabling a connection to be made.

This dataset contained a lot of repeated data and a substantial amount of errors and some

preparation was made to obtain clean data.
In table 3.1 there is a small data sample.

nº guia	transporte ID	Chegada fabrica	inicio descarga	fim descarga	saida fabrica	local de descarga	part number
000180576	219779693	03.01.19 10:03	03/01/2019 10:54:00	03/01/2019 11:30:00	03/01/2019 12:23:00	LOZ_5_KLT	6R0915105B
016648930	219812549	04.01.19 01:00	04/01/2019 01:35:00	04/01/2019 01:50:00	04/01/2019 01:51:00	LOZ10_GLT	1S0915105A
016648931	219812549	04.01.19 01:00	04/01/2019 01:35:00	04/01/2019 01:50:00	04/01/2019 01:51:00	LOZ10_GLT	5TA915105B
016648932	219812549	04.01.19 01:00	04/01/2019 01:35:00	04/01/2019 01:50:00	04/01/2019 01:51:00	LOZ10_GLT	7P0915105
016648927	219809991	04.01.19 01:36	04/01/2019 02:04:00	04/01/2019 02:17:00	04/01/2019 02:18:00	LOZ10_GLT	1S0915105A
016648928	219809991	04.01.19 01:36	04/01/2019 02:04:00	04/01/2019 02:17:00	04/01/2019 02:18:00	LOZ10_GLT	7P0915105
016648929	219809991	04.01.19 01:36	04/01/2019 02:04:00	04/01/2019 02:17:00	04/01/2019 02:18:00	LOZ10_GLT	5TA915105B
000180616	219843212	04.01.19 05:28	04/01/2019 05:54:00	04/01/2019 06:34:00	04/01/2019 06:35:00	LOZ_5_KLT	6R0915105B
016649484	219891419	07.01.19 07:05	07/01/2019 08:04:00	07/01/2019 08:37:00	07/01/2019 10:07:00	LOZ_5_KLT	7P0915105A

Table 3.1: External transport data sample

3.2 Receiving Data

Regarding the receiving cluster the data available consisted on excel files with records of material orders. The files contained entries for each unit load ordered and had the following information, material order creation date, truck license plate, truck arrival data, part identification number, supplier information and the warehouse position for the unit load. In this case the material order number allows connection to the external transport data and the warehouse position for the unit load allows connection to the warehousing data. On table 3.2 we can see a data sample of this data.

Area	Nr. Fornec	Nr Guia	Posição	Dt Guia	Dt Entrada	Peça	Gr Arm	Embalagem
FCC1	0001551600	000180576	MS05A04A03	2018-12-19	04/01/2019 00:30	6R0915105B	T2	DB0011
FCC1	0001551600	000180576	MS05A08A03	2018-12-19	04/01/2019 00:30	6R0915105B	T2	DB0011
FCC1	0001551600	000180576	MS05A09A01	2018-12-19	04/01/2019 00:30	6R0915105B	T2	DB0011
FCC1	0002522100	016648927	INSPECAO	2018-12-27	04/01/2019 03:35	1S0915105A	T2	DB0011
FCC1	0002522100	016648927	INSPECAO	2018-12-27	04/01/2019 03:35	1S0915105A	T2	DB0011
FCC1	0002522100	016648927	MS05B22A02	2018-12-27	04/01/2019 03:36	1S0915105A	T2	DB0011
FCC1	0002522100	016648927	MS05B22A03	2018-12-27	04/01/2019 03:36	1S0915105A	T2	DB0011
FCC1	0002522100	016648927	MS05B27A03	2018-12-27	04/01/2019 03:36	1S0915105A	T2	DB0011
FCC1	0002522100	016648927	MS10B25A02	2018-12-27	04/01/2019 03:36	1S0915105A	T2	DB0011

Table 3.2: Receiving data sample

3.3 Warehousing Data

The warehousing data available consisted of a daily report that included the quantity and identification of the material stored in each occupied position of the VWAE internal warehouses.
This report is generated with the arrival of material from the receiving cluster and the transports of material leaving the warehouse, not a real time picture of the state of the warehouse, this can be a problem because it can lead to data errors like for example if during the warehousing process an operator stores a container in a wrong position the data will show the container at the correct position.

Table 3.3 is a data sample of the described data.

Area	NrReferencia	Zona	Loc.	Peça	Fornecedor	QStatus	GrArm	Embalagem	Nr. Guia	Dt Guia	Ultimo Mov.	Quantid
43B1	04314028017754	PSO	BN05A14D01	7M3810630A	00156324	00X	B9	0015SCH	007011214	01/12/2014	13/07/2017	116
43B1	04316031877318	U20	BN06B04E01	7N0864633A	00153479	00X	K3	0006PAL	026301258	01/09/2016	13/03/2018	500
43B1	04317035155166	PSO	BN05A14B02	1K8827209A	00057588	280	B8	111902	000397189	03/03/2017	23/02/2018	28
43B1	04317036301736	PSO	BN05A14B03	1K8827210A	00057588	280	B8	111902	000405075	05/07/2017	23/02/2018	12
43B1	04317036919162	V05	BN03B03C01	1K0809495	00016954	280	B8	111902	000379297	14/09/2017	20/10/2017	500
43B1	04317036938926	INK	MAT-NOK.	1K8864629B	00051288	000	68	006280	002161428	04/09/2017	02/11/2017	200
43B1	04317036950013	ING	MAT-NOK	1K0813146	00071142	280	B8	111950	040125786	05/09/2017	13/02/2018	129
43B1	04317036968762	V05	BN03B02C02	1K0809495	00016954	280	B8	111902	000377602	04/09/2017	20/10/2017	361
43B1	04317036978923	BKL	BN99011A05	7N0864623A	00153479	00X	K2	0001SCH	060705981	08/09/2017	25/10/2017	1080
43B1	04317036978927	BKL	BN99011A05	7N0864623A	00153479	00X	K2	0001SCH	060705981	08/09/2017	25/10/2017	1080

Table 3.3: Stock data sample

3.4 Transport to Sequencing

When it comes to internal transports there is a system responsible for the movements of tow tugs and forklifts. This system manages all internal transports apart from AGV's, and stores the information regarding each internal transportation, creating a huge amount of data every day.

This data is very detailed and contains a description of each movement made, however the size of this data was overwhelming for the traditional tools for data analysis used like excel, because in a single day hundreds of thousands of records can be generated and weigh over 30 megabyte, that led to this data being left unexplored by the VWAE planners because they only have traditional tools like excel.

The first step made was to isolate the entries regarding transport of the car batteries to compare with the data from the other clusters and with the shop floor situation to understand the meaning of the data. Then a data validation step was made with specialists on these databases from the VWAE in order to understand the real meaning of the data, situations like sensor errors, or impossible values can be quickly detected and explained by specialists in the logistics process, this work with specialists was made for each data source but the complexity of this data led to this data being the focus of our meetings.

After this work with the specialists I advanced to clean and prepare data for all existent car components.

This data was connected with the warehouse and sequencing data using the partnumber and warehouse position fields.

ITLS-Auftrags-Nr.	Sachnummer Karte	DtHrMov_PT	Ereignisschlüssel	Verbindername	Subsystem	Menge	Lagerplatz
483737	3Q0813116B	08/01/2020 23:02:59	subsystem_started	BODY_P1E1->BODY_P1	SLS_BODY	20	B-MAKE
483737	3Q0813116B	08/01/2020 23:04:21	subsystem_finished	BODY_P1E1->BODY_P1	SLS_BODY	20	B-MAKE
565637	BUNDLE-R-DE	08/01/2020 23:04:40	subsystem_started	POT-KLT->BHF-KLT-MF	SLS_ASSY	28	
565637	BUNDLE-R-DE	08/01/2020 23:04:42	subsystem_finished	POT-KLT->BHF-KLT-MF	SLS_ASSY	28	
483743	2GA821105A	08/01/2020 23:04:59	subsystem_started	BODY_P1E1->BODY_P1	SLS_BODY	26	B-MAKE
483743	2GA821105A	08/01/2020 23:05:17	subsystem_finished	BODY_P1E1->BODY_P1	SLS_BODY	26	B-MAKE
483720	2GA809642	08/01/2020 23:05:22	subsystem_started	PAL1_A2_LPL->PAL1_EXPED.B2	SLS_PAL1	60	GESTAMP-01
483720	2GA809642	08/01/2020 23:05:28	subsystem_finished	PAL1_A2_LPL->PAL1_EXPED.B2	SLS_PAL1	60	GESTAMP-01
565637	BUNDLE-R-DE	08/01/2020 23:05:29	subsystem_started	BHF-KLT-MF->VBHF-KLT-DE	ZLS-KLT-MF	28	
43B1	04317036978927	BKL	BN99011A05	7N0864623A	00153479	00X	K2

Table 3.4: Internal transport data sample

3.5 Sequencing Data

The Sequencing data available consisted of the logs from the scanners used by the sequencing operators. One issue was that data was only available for consultation for short periods of time (15 days) and was not being stored. A few reasons explained why this data was not being stored, the data had a lot of noise, the logs available had multiple of data repeated, and multiple columns that had no value, and that made the data heavy. The process of accessing these logs was also a time-consuming task.

The first solution for our use case was the manual process of accessing the data for the car batteries parts and storing the data locally. This served the short-term goals but failed the requirements of scalability. For a more robust solution a Python script was created to access and download the data for the batteries case automatically using python libraries for web scrapping beautiful soup and selenium.

This script also contains operations of data cleaning, merging and storage autonomously. To a viable scale up for all parts root changes on this database and reporting system were necessary and suggested to the "owners" of the system within VWAE. These suggestions consisted on the removal of some redundant data and the inclusion of some key fields in the data and were well received and included in an update patch of the sequencing data system, now sequencing data for every car component is already being gathered and prepared automatically.

Table 3.5 is a data sample of this data.

DATE	VALUE1	VALUE2	VALUE3
2019-03-19 23:59:58:1167009	Info	PICK	Put Position 3 OK (Expected 3)
2019-03-19 23:59:58:1167009	Info	PICK	Put OK
2019-03-19 23:59:58:1011008	Bluetooth	PICK	3
2019-03-19 23:59:58:1011008	Info	PICK	Verify Put
2019-03-19 23:59:51:9558903	Info	PICK	Pick OK
2019-03-19 23:59:51:9402903	Bluetooth	PICK	PN3Q0825236D
2019-03-19 23:59:51:9402903	Info	PICK	Verify Pick PN3Q0825236D(Expected 3Q0825236D)
2019-03-19 23:59:51:9402903	Info	PICK	Verify if BC contains value(s) 3Q0825236D
2019-03-19 23:59:47:1990823	Info	PICK	Background Color WHITE

Table 3.5: Sequencing data sample

3.6 Transport to POF

Regarding the transport from the sequencing areas to the point of application is made either by AGV's or again tow trucks. AGV's have their data internal data (logs and sensor) stored in different databases depending on its manufacturer.

The data from the processes with tow tugs was already described as the process is the same from transports from the warehouse to the sequencing area.

For our use-case the car batteries were transported from the sequencing area to the point of fit by an AGV. However, this AGV belongs to one of the oldest generations of AGV's

present in VWAE and has no type of connection to any database. Regarding this transportation there was no direct data available.
To counter this issue, we prepared a raspberry pi with movement and position sensors and attached it to the AGV for some basic data gathering about the AGV behaviour and workload.
This step was also made to ensure data validity and plausibility, by doing it we were able to compare the transport times measure with the times recorded by the sequencing cluster and by the production line. This data validation is very important for planners, if the data is correct the KPI's (Key Performance Indicators) obtained from it can be relied upon.

3.7 POF Data

Point of fit data consist of logs from the production line. Each car as a tag that communicates with receivers distributed along the production line, this attributes a time stamp. The files available are the logs from the receivers that contain the information about each car with a timestamp associated.
From this data I can for example draw a graph of the number of cars produced each day at VWAE. The first 100 days of 2018 production at VWAE are represented in figure 3.1. We can see that the normal production is steady around 900 cars a day and there and the weekends with zero production. Also, a notable stoppage of production in the end of the month of march is explained by the holidays.

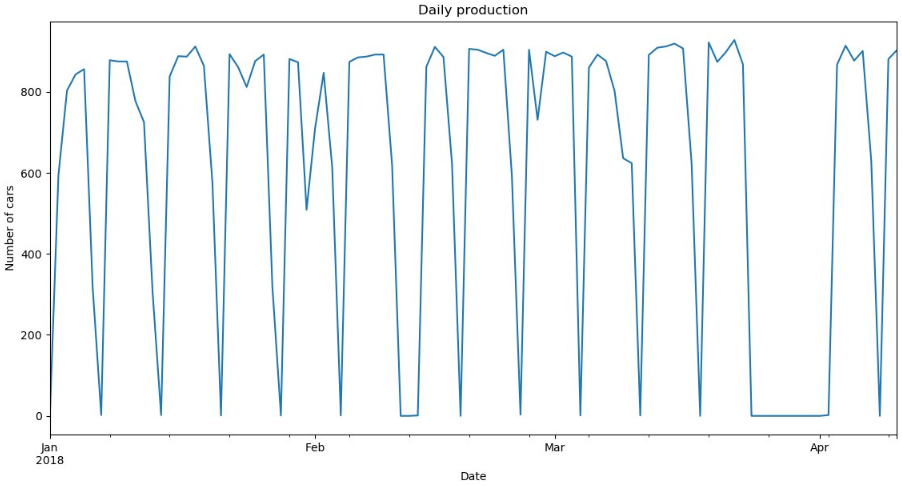


Figure 3.1: Daily production first 100 days of 2018

For the purposes of our use-case, where car batteries are the part selected, we consulted the log files from the receiver placed at the point of application of the batteries on the car in the production line. This data comes in excel format (.xlsx) and in a tabular shape, with 5 columns, one with the timestamp, other with the car identification number, and the other tree with info about the car and the number of entries is equal to the number of cars that passed that receiver during the time frame observed.

The totality of the dataset gathered during this study consists of the logs from the receiver correspondent to the car batteries for the start of 2018 to the end of February 2020, this dataset will be used to predict the future consumption by the production line of each type of batterie further on this thesis.

Table 3.6 is a data sample of the POF data.

KNR	**Sequence**	**Date - T300**	**FAM BAT**	**FAM AAU**	**FAM MOT**	**FAM GSP**	**Model**
5240356	8602	02/01/2018 07:12	J0V	E0A	D60	G1A	7N2
5140145	8603	02/01/2018 07:13	J0T	E0A	DQ6	G1D	A11
5110392	8604	02/01/2018 07:15	J0V	E0A	DN4	G1D	A11
4930177	8605	02/01/2018 07:16	J0S	E0A	DS9	G0K	A11
5250312	8606	02/01/2018 07:22	J0V	E0A	D60	G1A	7N2
4930189	8607	02/01/2018 07:24	J0S	E0A	DS9	G0K	A11
4930337	8608	02/01/2018 07:25	J0S	E0A	DS9	G0K	A11
4930015	8609	02/01/2018 07:26	J0S	E0A	DS9	G0K	A11
5250314	8610	02/01/2018 07:29	J0V	E0A	D60	G1A	7N2

Table 3.6: POF data sample

3.8 Inventory Data Analysis

This section will include a summary of a meeting with the planners responsible for the car batteries at VWAE and a description of the inventory data.

This meeting was scheduled to understand the biggest challenges found by the planners through the years understand and their day to day tasks as well as some tasks of data validation.

The planners started by pointing out the complexity of the logistics involved in a car factory like VWAE, there are hundreds of suppliers from all over the world, thousands of different components and not all of them go through the same processes, for example a component produced in the industrial park next to VWAE is treated differently from a component produced in Poland. Each planner is responsible for a list of car components and although data from all clusters for each component exist accessing it is time consuming and requires the utilization of multiple systems and platforms.

One of the questions asked in this meeting was "How long does it take from a order being placed and the material reaching VWAE?" and the response was that it varies a lot depending on the component and supplier, and for the specific case of car batteries it is usually around 4 or 5 days. They also pointed out that material transports always depend on external and sometimes unpredictable factors such truck drivers' strike, transports infrastructures or even weather conditions and when these situations occur, they need to react quickly to avoid shortages.

The inventory data was generated with the purpose of analysing the problem in study is based on the difference between entries of batteries to the warehouse (receiving data) and the supplying of batteries to the sequencing area (internal transport data), and for the initial state of the warehouse levels I utilized the warehousing data for the first day of production in 2018.

This dataset has data of the entire 2018 year and the first 3 months of 2019.

The data consists of a table with 6 columns and 12861 rows, the first column being the date and hour of the day and the following 5 the number of stored packages for each battery type.

Since this dataset reflects the magnitude of the problem at hand an analysis was made.

The first step of this was to define what would be an optimal value, car batteries were selected to this study because each and every car that is produced in VWAE utilizes one and only one batterie, but there are still multiple batterie types, in this particular case five of them, each one with a different usage rate by the production line which is called take rate.

Based on the take rate we create two categories on the batterie types, low runners for types with have a take rate below 10% and high runners for those above.

Since a stop in production is very expensive that cannot happen because of material shortage situations and to unsure that management and specialists from the logistics department at VWAE calculate, based on the importance of the part and on supplier localization, a security stock for each part, and the inventory levels should never dip below that level, except on shutdown occasions.

Based on feedback from the planners responsible for these components a security stock level of two and a half days was considered for all batterie types present.

To establish a baseline of overstock a steady production of 900 cars a day was considered, so we will consider that at the start of each day, we should have an inventory level of 3.5 days, the production of the day itself plus the security stock level of 2.5 days.

For each batterie type this baseline was obtained by multiplying the daily production by the take rate and then dividing that number by the number of batteries per container. This way we obtain the daily consumption in packages for each type of batteries.

As an example, let us consider the 1S0915105A batterie type. ((900 * 0,4383) 54) = 7,3 packages per day but since we only store full packages, we need to ensure 8 packages per day in this case.

Now all that there is left to do is multiply that value by 3.5 to get our reference value for this part. Another reference value we considered was of one week of production so again 8*7 = 56 packages.

The table 3.7 will have the results of this exercise for each type.

This values can be compared with the statistical indicators of the inventory levels, such as mean, standard deviation, maximum, minimum and percentiles of 25, 50, 75 and 90 to perceive the size of the problem.

The first indicator that situations of overstock occur frequently is that the mean value

Bat Type	7P0915105A	7P0915105	6R0915105B	1S0915105A	5TA915105B
mean	7,73	2,86	15,30	33,25	33,92
std	6,01	1,57	6,49	13,19	11,54
min	1,00	0,00	4,00	5,00	5,00
25%	4,00	2,00	11,00	22,00	25,00
50%	6,00	3,00	15,00	33,00	35,00
75%	8,00	4,00	19,00	43,00	42,00
90%	10,00	5,00	24,00	50,00	50,00
max	33,00	8,00	38,00	75,00	65,00
Take Rate 2018 %	2,52%	3,18%	13,42%	43,83%	37,05%
Units by package	36,00	48,00	48,00	54,00	48,00
Round up Units per day	23,00	29,00	121,00	395,00	334,00
Packages per day	0,64	0,60	2,52	7,31	6,96
Round up Packages per day	1,00	1,00	3,00	8,00	7,00
3.5 days	3,50	3,50	10,50	28,00	24,50
7 days	7	7	21	56	49

Table 3.7: Inventory analysis VWAE 2018

is bigger than our reference of 3.5 days for all types except for one that happens to be one of the low runners, and the maximum value is bigger than the 7 days usage for all the different types.

In the case of the high runners we observe that our 3.5-day reference is always smaller than the 75th percentile so we can safely say that in at least 25% of the time we are facing situations of overstock, and still on the case of the high runners two of them have inventory levels superior to 7 days of consumption 10% of the time.

To illustrate this the graphs of the inventory levels for the month of September of 2018 of the high runners 1S0915105A and 5TA915105B are presented in the figures 3.1 and 3.2. In both graphs we have the inventory levels and the reference line is plotted with the value of the 3.5 days of production as explained before and represented on table 3.1.

Both of these have a great area above the reference lines and some observations can be made.

When the inventory levels go up in a given instant it means that a truck with new batteries was received at the factory. There are multiple occasions on both of these graphs where situations of overstock were already happening, and a new batch of material was unloaded. In 2018 more than 1770 containers corresponding to over 95 thousand batteries distributed in over 100 trucks of the high runner 1S0915105A were unloaded at VWAE. This goes to show that there is room for improvement that can bring multiple optimizations on this process alone, like reductions of the number of trucks (Co2 and money), reduction of inventory space occupied and reduction of stall money. If situations like these are detected in advance by our system, it can alert inventory management specialists and suggest improvements.

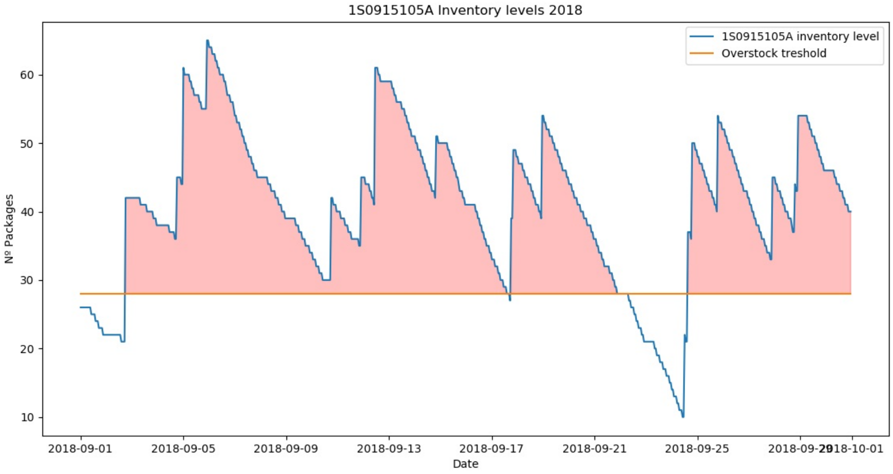


Figure 3.2: Inventory levels September 2018 1S0915105A

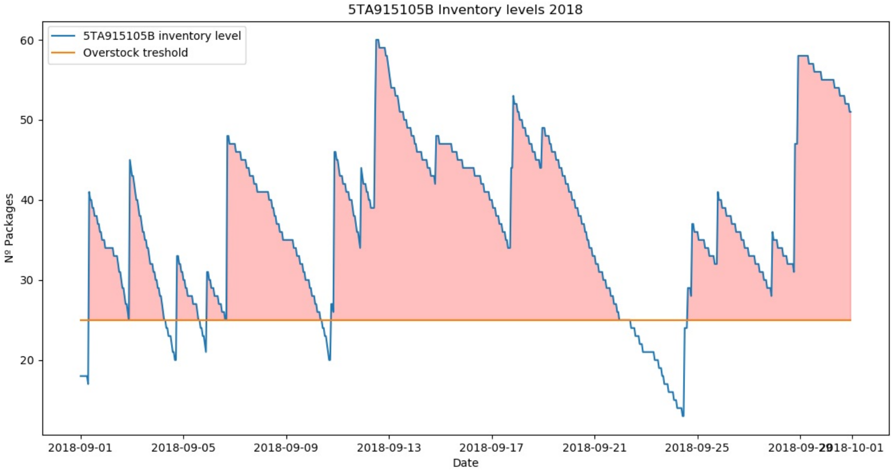


Figure 3.3: Inventory levels September 2018 5TA915105B

Chapter 4

Architecture

This chapter will describe the architecture implemented in the proposed solution to solve the problem presented before.

The objective is to implement a data-driven system capable of improve efficiency in the intralogistics processes at VWAE and to that end a layered architecture was chosen to maintain flexibility and scalability as layers allow to test and work on components independently of each other, changes to one of the layers do not require changes in others, the usage of layers helps to control and encapsulate the complexity of large applications and with a layered approach multiple applications can effortlessly reuse the components. Since the objective is a data-driven system the first layer is the ETL layer that will gather, prepare and load all the data available to the storage layer. This layer is then connected to a machine learning layer or directly to a processing layer that processes the output of the machine learning layer and connects to our visualization layer that presents the logistics data in a visual way to planners.

On figure 4.1 we can see the different layers and the flow of data from the collection on the shop floor to the data visualization layer.

4.1 Extract Transform Load layer

In order to utilize the gathered data in a fruitful way and to apply machine-learning algorithms data needs preparation. This layer is responsible for getting the data gathered and transform it in a format acceptable by the storage layer and by the machine learning layer. Each cluster of data has data in different formats and different sources has we observed in the logistics data analysis chapter of this document, each of those clusters required different operations of ETL and some of the more important are described in the logistics data analysis chapter of this document.

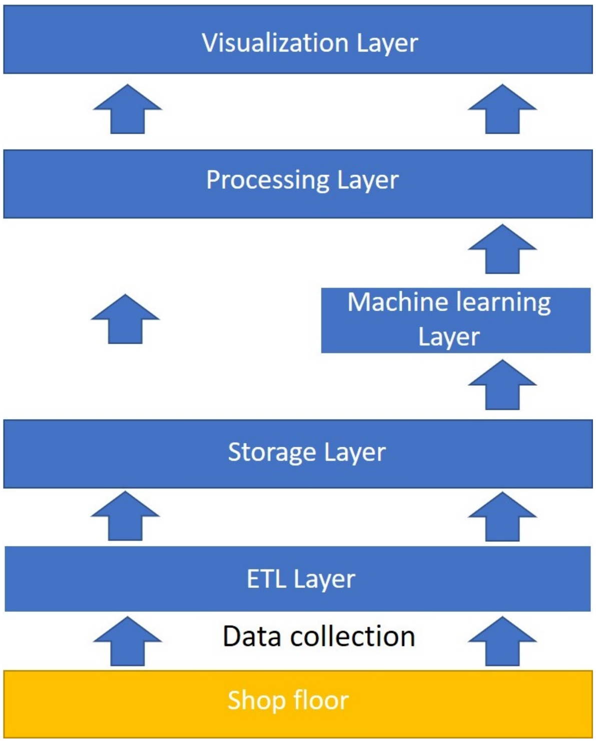


Figure 4.1: Architecture overview

The different sources have different processes to gather data nonetheless all of the different data is gathered in excel format. This layer cleans and prepares the data using python scripts to then loads it to the storage layer with a python connection.
Our machine learning layer needs data of the production of batteries dataset in a specific format to predict the future production cars with that batterie, the steps to prepare the data were made with the Python library pandas[65] [66] and will be described here.
The raw data we receive from the production systems consist of a excel table that has the 5 following columns, car identification number, sequence number, date, type of car batteries and model of the car, and we have records for the entire 2018 and 2019 years.
After a quick analysis to this data we can see that it's not prepared for machine learning input the first step was dropping the sequence number, because it served no purpose for our machine learning objectives.
The machine learning layer requires data with a given frequency or time steps, since we have production data and we have no time step defined, each entry represents a car produced at a given moment, to solve this issue, we decided to resample the data to a daily format and created a dataset with the index being the dates from the first day of 2018 to the last day of 2019. After this operation our dataset now has one entry for each day and has 8 columns, one for each batterie type(5) and one for each model(3) produced, this is exemplified in table 4.1.

After this some operations on the data were made in a trial and error based on the performance results from the machine learning layer.

Date	J0S	J0T	J0V	J1N	J2D	711	7N2	A11	Month	weekday	Year	Week
01/01/2018	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1	0	2018	1
02/01/2018	269.0	150.0	136.0	23.0	16.0	38.0	106.0	450.0	1	1	2018	1
03/01/2018	349.0	36.0	364.0	19.0	35.0	60.0	134.0	609.0	1	2	2018	1
04/01/2018	183.0	49.0	550.0	21.0	40.0	81.0	122.0	640.0	1	3	2018	1
05/01/2018	519.0	31.0	263.0	14.0	29.0	73.0	134.0	649.0	1	4	2018	1
06/01/2018	184.0	39.0	83.0	8.0	4.0	29.0	48.0	241.0	1	5	2018	1
07/01/2018	2.0	0.0	0.0	0.0	0.0	0.0	0.0	2.0	1	6	2018	1
08/01/2018	583.0	35.0	220.0	20.0	20.0	77.0	135.0	666.0	1	0	2018	2

Table 4.1: Input Data

The model with better results in terms of accuracy had as additional columns with the month number of the date, information about day of the week, and the calendar week number. These were added with the hope that some of these basic additions can make some patterns more evident to the machine learning model and we got positive results.
One other test that was made but did not improve accuracy significantly was the introduction of lag values, each entry of data would have a set of columns with the value from the n entries before. As this operation did not prove advantageous it was dropped.
After these steps a simple division on training and test datasets was made.
Finally, before being feed into the machine learning model the data needs to normalize to numbers between 0 and 1. To achieve this we utilized the MinMaxScaler feature from the Scikit library.
Scikit-learn[67] is an open source machine learning library that supports supervised and unsupervised learning. It also provides various tools for model fitting, data preprocessing, model selection and evaluation, and many other utilities.

4.2 Storage layer

This layer consists database with all the data described on the data analysis chapter. This layer is important to store data continuously and effortlessly and to eliminate the loss of historical data, this way the growing amount of data can feed the analytics and improve their value.
For this layer multiple databases were considered both relational and non-relational. Non-relational databases offer multiple data models, are easily scalable and can be faster that relational databases however some of them are incompatible with the ACID (atomicity, consistency, isolation, durability) properties, relational databases like PostgreSQL [68] offer ACID compliance, ease of both implementation and maintenance as well as a standard query language, but as a drawback relational databases can be difficult to scale[69] [70].
During the data understanding and data preparation phases of this dissertation became evident that the data from the VWAE logistics processes is highly related between sources therefore a relational database was chosen in specific PostgreSQL because it is one of the more popular open-sourced solutions available.
Figure 4.2 shows the diagram of this database with the different tables and connections

made between them. This database was created to store all of the historical data described in the chapter 3 of this document. The objects of this database and connections between them were created in a way to integrate all of the part numbers present at VWAE and this process wouldn't be possible without the data analysis made before, it was this process that identified the connections between the different data clusters, and the valuable and redundant information.
Data is now structured, clean and integrated in a single database.

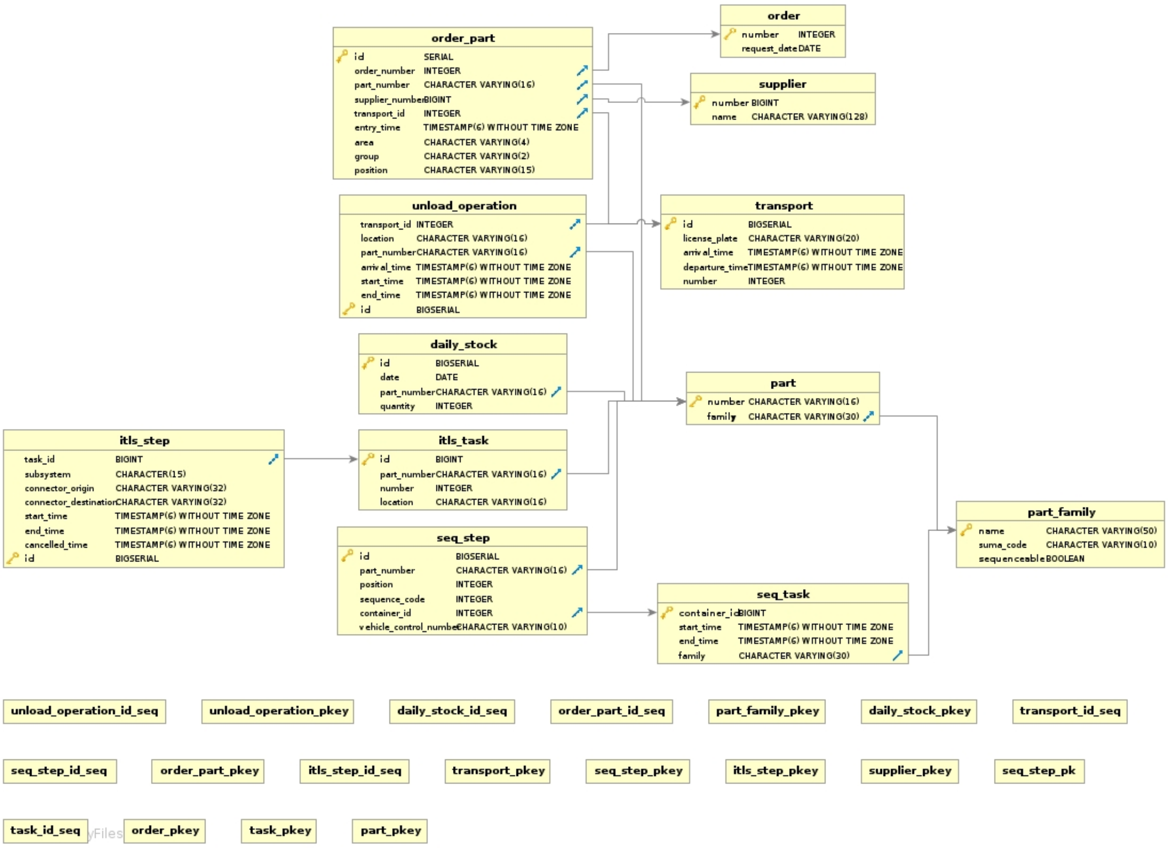

order
number INTEGER
request_date DATE
order_part
id SERIAL
order_number INTEGER
part_number CHARACTER VARYING(16)
supplier_number BIGINT
transport_id INTEGER
entry_time TIMESTAMP(6) WITHOUT TIME ZONE
area CHARACTER VARYING(4)
group CHARACTER VARYING(2)
position CHARACTER VARYING(15)
supplier
number BIGINT
name CHARACTER VARYING(128)
unload_operation
transport_id INTEGER
location CHARACTER VARYING(16)
part_number CHARACTER VARYING(16)
arrival_time TIMESTAMP(6) WITHOUT TIME ZONE
start_time TIMESTAMP(6) WITHOUT TIME ZONE
end_time TIMESTAMP(6) WITHOUT TIME ZONE
id BIGSERIAL
transport
id BIGSERIAL
license_plate CHARACTER VARYING(20)
arrival_time TIMESTAMP(6) WITHOUT TIME ZONE
departure_time TIMESTAMP(6) WITHOUT TIME ZONE
number INTEGER
daily_stock
id BIGSERIAL
date DATE
part_number CHARACTER VARYING(16)
quantity INTEGER
part
number CHARACTER VARYING(16)
family CHARACTER VARYING(30)
itls_step
task_id BIGINT
subsystem CHARACTER(15)
connector_origin CHARACTER VARYING(32)
connector_destination CHARACTER VARYING(32)
start_time TIMESTAMP(6) WITHOUT TIME ZONE
end_time TIMESTAMP(6) WITHOUT TIME ZONE
cancelled_time TIMESTAMP(6) WITHOUT TIME ZONE
id BIGSERIAL
itls_task
id BIGINT
part_number CHARACTER VARYING(16)
number INTEGER
location CHARACTER VARYING(16)
part_family
name CHARACTER VARYING(50)
suma_code CHARACTER VARYING(10)
sequenceable BOOLEAN
seq_step
id BIGSERIAL
part_number CHARACTER VARYING(16)
position INTEGER
sequence_code INTEGER
container_id INTEGER
vehicle_control_number CHARACTER VARYING(10)
seq_task
container_id BIGINT
start_time TIMESTAMP(6) WITHOUT TIME ZONE
end_time TIMESTAMP(6) WITHOUT TIME ZONE
family CHARACTER VARYING(30)
unload_operation_id_seq
unload_operation_pkey
daily_stock_id_seq
order_part_id_seq
part_family_pkey
daily_stock_pkey
transport_id_seq
seq_step_id_seq
order_part_pkey
itls_step_id_seq
transport_pkey
seq_step_pkey
itls_step_pkey
supplier_pkey
seq_step_pk
task_id_seq
yFiles
order_pkey
task_pkey
part_pkey

Figure 4.2: DB diagram

4.3 Machine learning layer

For our system to work a prediction of the usage of car batteries on the VWAE production line was necessary. To this end, we recurred to machine learning models. Multiple models were experimented with different parameters and features and the chosen model was Long short-term memory (LSTM) which is an artificial recurrent neural network architecture. This choice was made because LSTM are reportedly very good at forecasting time series data and do not require a lot of parameterization for multivariate datasets. To implement this model, we used an already built solution for LSTM in python from the

KERAS API[38].

Since in the previous layer (ETL layer) we prepared the data for machine learning input all we need to do is create and feed a model. The objective is to input all of the production data available at the moment for each type of batteries and predict production for the next 3 days, to allow us to calculate the exact amount of batterie orders we need to place. To simplify we decided to predict the production of a single type of batterie the 1S0915105A(J0S), however some tests to predict the production of the other high runners showed similar results as we could expect.

For evaluation purposes we decided to use cross validation which is a technique to evaluate the performance of ML models, the objective of cross validation is to test the model's ability to predict new data that was not used in estimating it, in order to flag problems like overfitting or selection bias and to give an insight on how the model will generalize to an independent dataset, in this case we divided the input dataset in 90% of the records for training of the model and the last 10% for testing the model, and also the model will set apart 10% of the training data, will not train on it, and will evaluate the loss on this data at the end of each epoch. The loss function that we are utilizing is the MSE (Mean Squared Error), that like the name says measures the average squared difference between the estimated values and the actual value.

The process of building a machine learning model is iterative and throughout this process various combinations of models and parameters have been tested and the choice of the implemented model and its parameters was based on the performance of the different models tested.

One of the first decision was the steps ahead that our model would forecast, in this case we are utilizing daily data which means that each time step corresponds to one day, since we need to predict 3 days in to the future this parameter was locked on 3. This means that our model will try to predict 3 days after the last data provided.

Throughout this process it became clear that for our data and models we need at least 50 epochs of training because the validation and training errors would consistently drop during the first 50 epochs and that more than 200 epochs of training are impractical since from this point most of the models showed no significant improvements in performance and in some cases the validation error climbed which is a sign that the model is overfitting.

Regarding optimizers, which are algorithms or methods used to change the attributes of your neural network such as weights and learning rate in order to reduce the losses, we experimented with "Adam" and "RMSProp" and ended up using "Adam" because it we got better results.

For the batch size we tried multiple values and ended up choosing 64 because it was the one with better results without compromising the training time.

Regarding the optimal number of layers, we observed that one and two layers of LSTM presented similar results, but the addition of more layers would result in worse performances.

To prevent our model from overfitting we inserted a dropout a layer. Dropout is a technique for addressing this problem. The idea is to randomly drop units (along with their connections) from the neural network during training. This prevents units from co-adapting too much and improves the performance of neural networks in multiple tasks [71].
The model implemented consisted of one LSTM layer with 2000 units and a dropout of 20%, we utilized a batch size of 64, and trained for 200 epochs. The model was compiled using the “Adam” optimizer and the training and validation errors evolution across epochs are shown in the results chapter in figure 5.5 and the predictions for the last 15 days in green as well as the true values in red are shown in figure 5.6.

4.4 Processing layer

In this section some calculations are needed to provide the optimizations on the inventory levels and order placement.
Here we take the predictions outputted by our machine learning layer to establish the necessary material for the desired time range.
Then with that information we consult the hourly stock to evaluate if the inhouse stock can meet the production for that time range.
For each batterie type the system will perform a simple subtraction, the available stock at the moment minus the production for the desired time range. We also need to subtract the security stock because we want to avoid dropping below this level. If this calculation returns a positive number, it means that we are facing a situation where the available stock will meet the demands without the need for new material. If it’s a negative number it means we ´ re facing a situation where the stock will not meet the demands for the entire time range without new material and the absolute value of this number is the quantity of new material that is necessary to meet the demand. Obviously if this number is zero, we are facing a limit situation where the available stock is even with the demand.
Note that at this point our intention is not to provide the optimal stock level for each part, but it’s to provide optimizations to a complex problem and to iteratively improve the system. By doing this we can minimize situations of overstock and reduce de number of warehouse space allocated to this part.
We intended for this calculations layer to be simple and objective as possible as the information gathered from data is already close to provide insights.

4.5 Visualization layer

Data visualization is the act of taking information (data) and placing it into a visual context, such as a map or graph.
Data visualizations make big and small data easier for the human brain to understand,

and visualization makes it easier to detect patterns, trends, and outliers in groups of data. Good data visualizations should place meaning into complicated datasets so that their message is clear and concise. We are an inherently visual world, where images speak louder than words. Data visualization is especially important when it comes to big data and data analyzation projects.

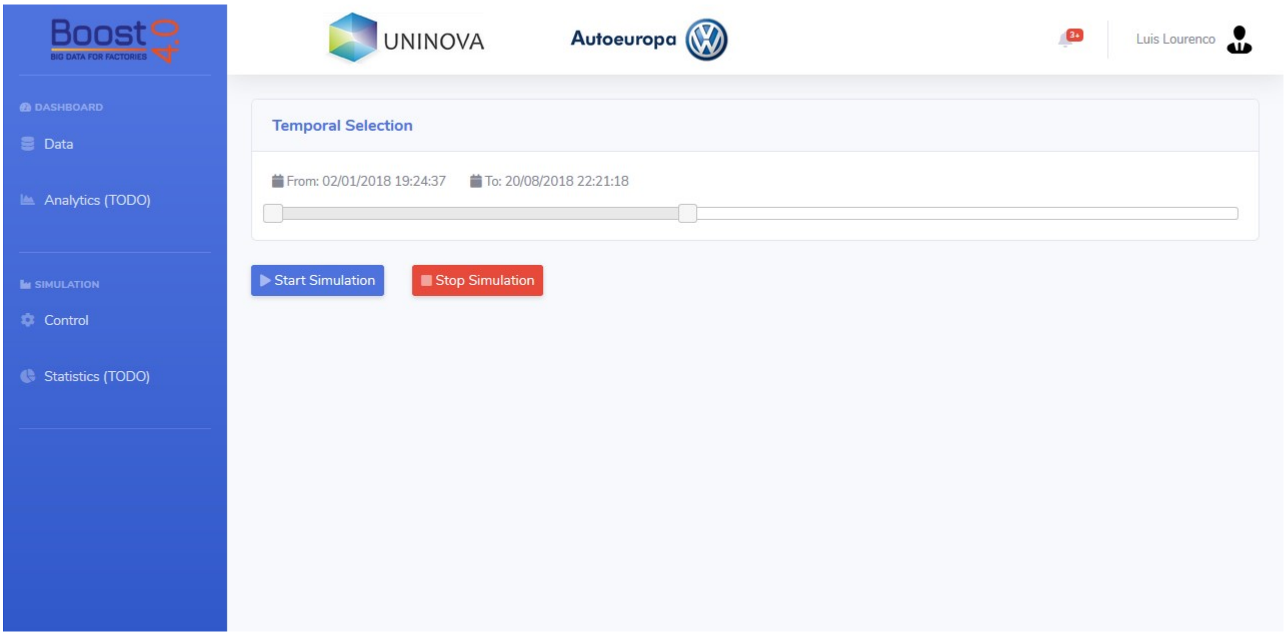
With this in mind and in order to get the most out of our data all of the features previously developed were aggregated and displayed in an interactive dashboard.

To build this dashboard we utilized an open source visualization and analytics software called Grafana. It provides charts, graphs, and alerts for the web when connected to supported data sources. It is expandable through a plug-in system. End users can create complex monitoring dashboards using interactive query builders.

This dashboard contains data from all of the analysed clusters with pre-defined views and graphs but also allows user interaction, like the ability to adjust the time window selected and apply filters to the data.

Another feature present is the automatic connection to the digital twin, this is visible in figure 4.3, where planners can select a time range and press the start simulation button and automatically an instance of the simulation software (Visual Components) starts with the selected data to allow the user to validate and gather insights from a simulation point of view with minimal effort.

In the figures 4.4 and 4.5 the temporal selection provides an easy way to visualize trends

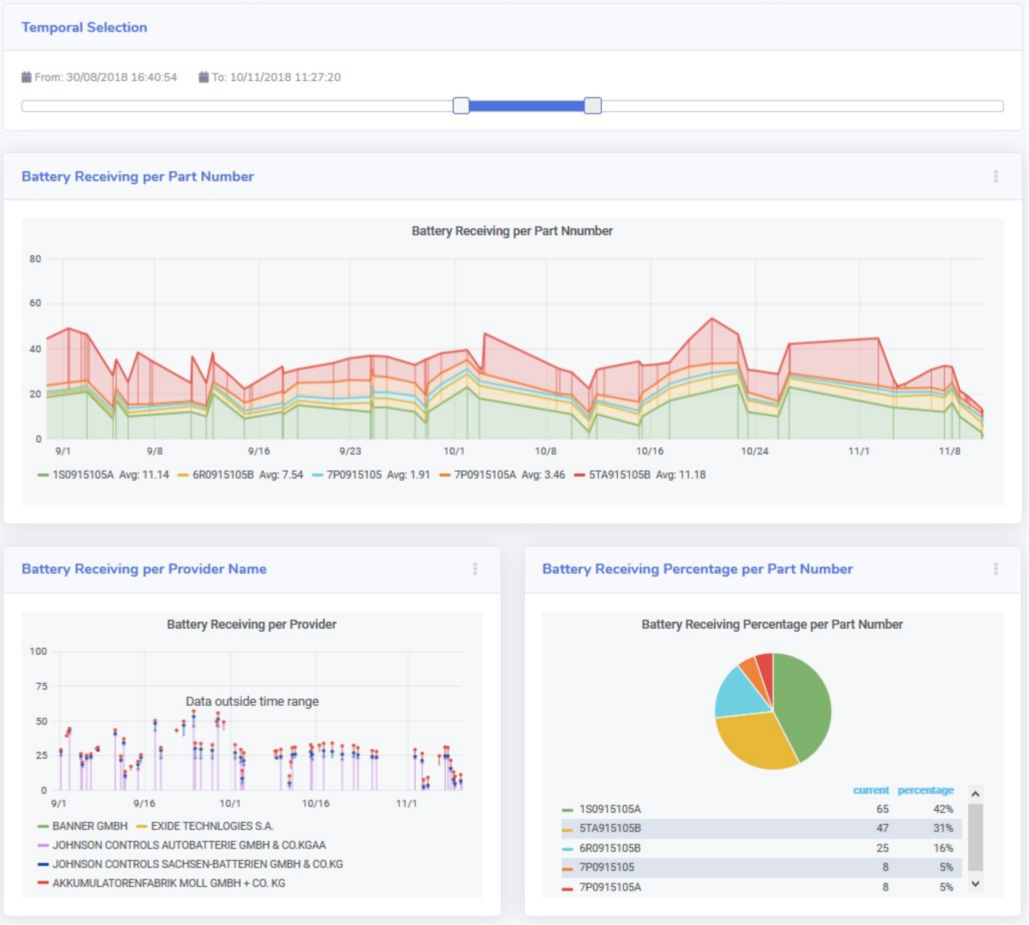


Temporal Selection

From: 02/01/2018 19:24:37 To: 20/08/2018 22:21:18

Figure 4.3: Digital Twin connection

and changes over time and across all clusters. Each cluster has a dedicated view in the dashboard where multiple graphics and tables are presented to users, in figure 4.4 one of the graphs represents the take rate of the five car batteries part numbers in VWAE and in figure 4.5 all 3 graphics represent the internal movements of containers regarding car



Temporal Selection

From: 30/08/2018 16:40:54 To: 10/11/2018 11:27:20

Battery Receiving per Part Number

Battery Receiving per Part Nnumber

Battery Receiving per Provider Name

Battery Receiving per Provider

100
75
50
25
0
Data outside time range
9/1
9/16
10/1
10/16
11/1
BANNER GMBH
EXIDE TECHNLOGIES S.A.
JOHNSON CONTROLS AUTOBATTERIE GMBH & CO.KGAA
JOHNSON CONTROLS SACHSEN-BATTERIEN GMBH & CO.KG
AKKUMULATORENFABRIK MOLL GMBH + CO. KG

Battery Receiving Percentage per Part Number

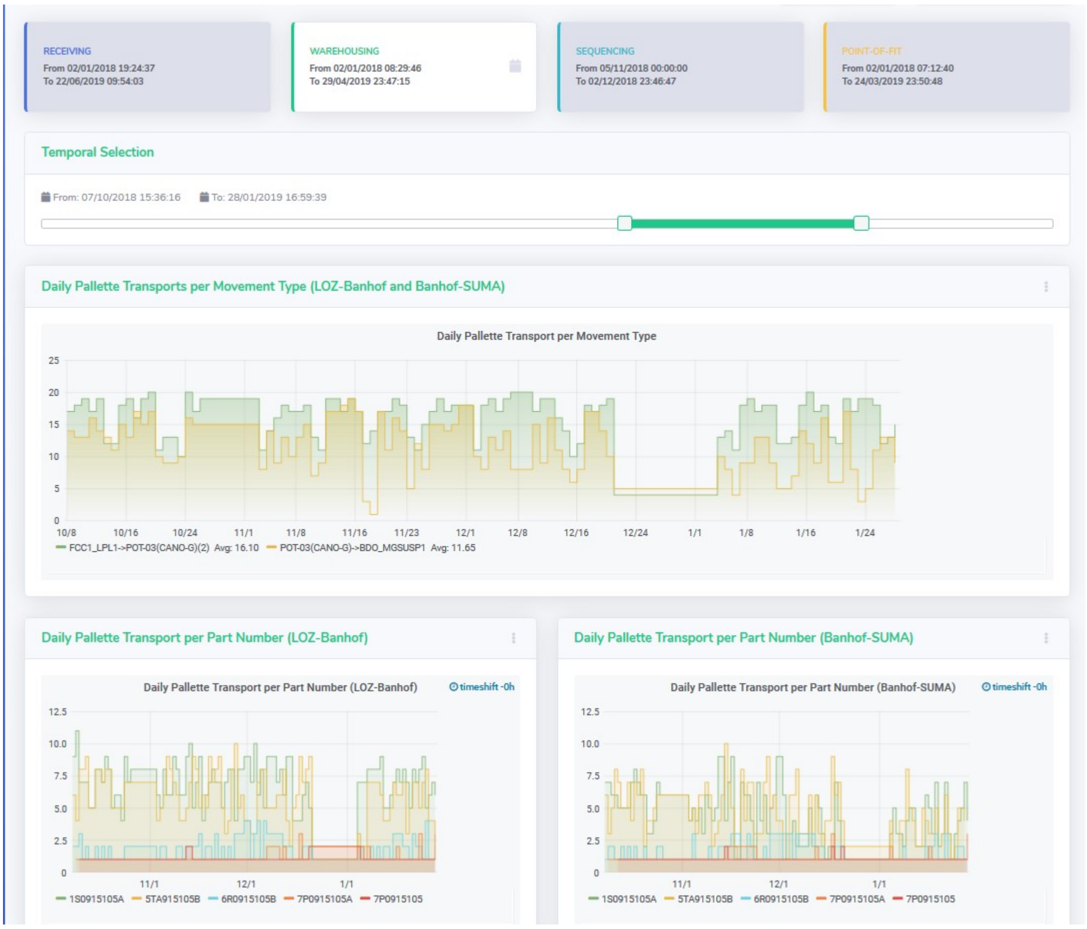
Battery Receiving Percentage per Part Number

	current	percentage
1S0915105A	65	42%
5TA915105B	47	31%
6R0915105B	25	16%
7P0915105	8	5%
7P0915105A	8	5%

Figure 4.4: receiving dashboard

batteries in the warehouse. The integration of datasets is visible in the figures and with this integration and visualization tools we eliminated all data-silos present in the logistics data available in VWAE. Planners can now in an intuitively manner consult data from multiple sources at the same time and also analyse historical data to make validations or look for patterns.

This functionality will be especially important when the planners are looking for make changes or even validate patterns recognized by the machine learning functionalities. It's all about providing information in an easy to read format to planners who will make the informed decisions.



RECEIVING
From 02/01/2018 19:24:37
To 22/06/2019 09:54:03

WAREHOUSING
From 02/01/2018 08:29:46
To 29/04/2019 23:47:15

SEQUENCING
From 05/11/2018 00:00:00
To 02/12/2018 23:46:47

POINT-OF-FIT
From 02/01/2018 07:12:40
To 24/03/2019 23:50:48

Temporal Selection

From: 07/10/2018 15:36:16
To: 28/01/2019 16:59:39

Daily Pallette Transports per Movement Type (LOZ-Banhof and Banhof-SUMA)

Daily Pallette Transport per Movement Type

25
20
15
10
5
0
10/8
10/16
10/24
11/1
11/8
11/16
11/23
12/1
12/8
12/16
12/24
1/1
1/8
1/16
1/24
FCC1_LPL1->POT-03(CANO-G)(2) Avg: 16.10
POT-03(CANO-G)->BDO_MGSUSP1 Avg: 11.65

Daily Pallette Transport per Part Number (LOZ-Banhof)

Daily Pallette Transport per Part Number (LOZ-Banhof)

Daily Pallette Transport per Part Number (Banhof-SUMA)

Daily Pallette Transport per Part Number (Banhof-SUMA)

Figure 4.5: warehousing dashboard

Chapter 5

Results

The objective of this chapter is to evaluate the impact of our system in optimizing the inventory levels at VWAE, our study focused particularly on car batteries and because of that all the results will be regarding this parts, however our approach can be applied to many of the car components existent at VWAE. The first part of this chapter presents the results of 3 machine learning models predicting the application of the 1S0915105A(J0S) batterie production and one for predicting the 5TA915105B(J0V) batterie to explain the decisions made throughout the process of building a machine learning model described in the architecture chapter of this document.

As stated in the architecture chapter throughout the process of building our machine learning model we provided different input data to the model, but to allow comparisons between the models presented all had the same input. This input data consisted of the normalized daily production data, and the addition of the weekday number and the calendar week number as showed in table 4.1.

Model I consisted on one LSTM layer with 2000 units and a dropout of 20% , we utilized a batch size of 64, and trained for 500 epochs. The model was compiled using the “RMSprop” optimizer and on figure 5.1 the training and validation errors (Y axis) evolution across epochs (X axis) are shown and on figure 5.2 the predictions of production of cars with the 1S0915105A batterie in the last 62 days of 2019 days in green as well as the true values in red are shown. With the validation error stabilizing with values very close to zero very different from the high validation error we can see that the model was overfitted.

Model II consisted on one LSTM layer with 3000 units and a dropout of 20% , we utilized a batch size of 64, and trained for 200 epochs. The model was compiled using the “Adam” optimizer and on figure 5.3 the training and validation errors (Y axis) evolution across epochs (X axis) are shown and on figure 5.4 the predictions of production of cars with the

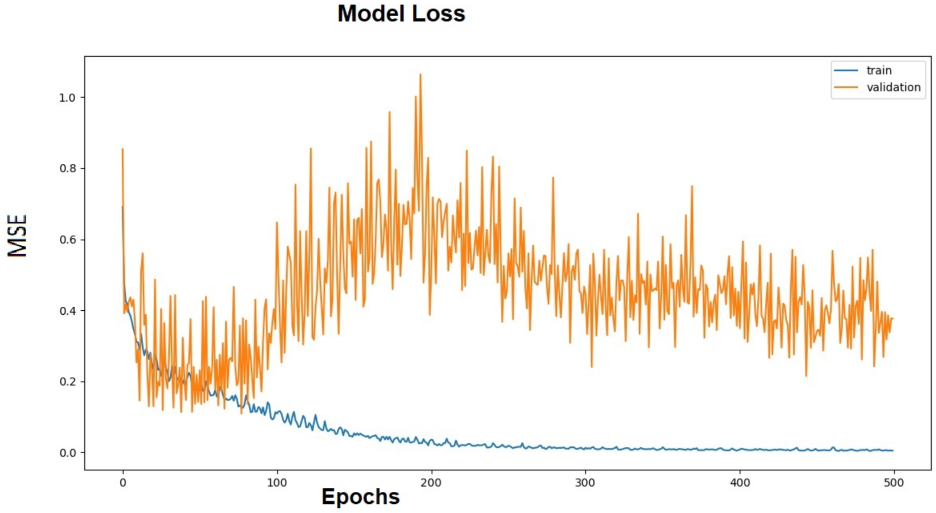


Figure 5.1: Error model I1

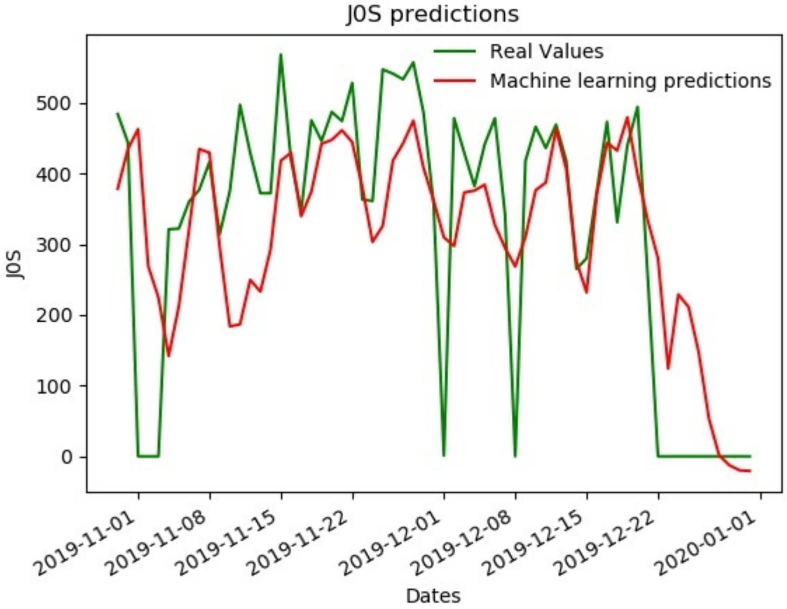


Figure 5.2: Predictions model I

1S0915105A batterie in the last 62 days of 2019 days in green as well as the true values in red are shown.

Model III consisted on one LSTM layer with 2000 units and a dropout of 20% , we utilized a batch size of 64, and trained for 100 epochs. The model was compiled using the "Adam" optimizer and on figure 5.5 the training and validation errors (Y axis) evolution across epochs (X axis) are shown and on figure 5.6 the predictions of production of cars with the 1S0915105A batterie in the last 62 days of 2019 days in green as well as the true values in red are shown. This was the model that presented better results therefore was

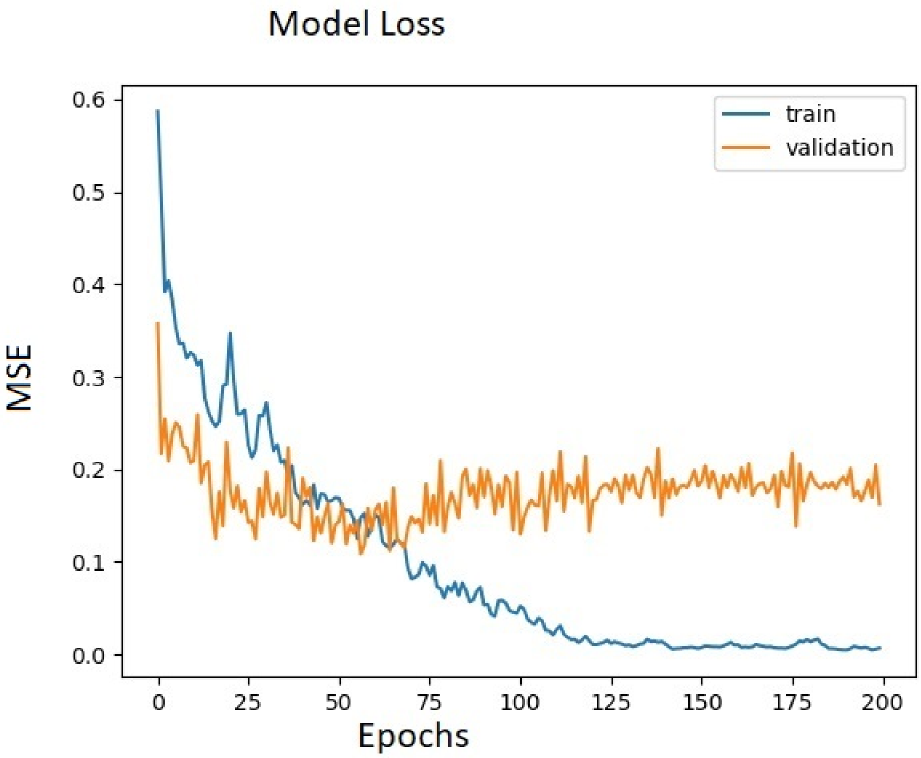


Figure 5.3: Error model II

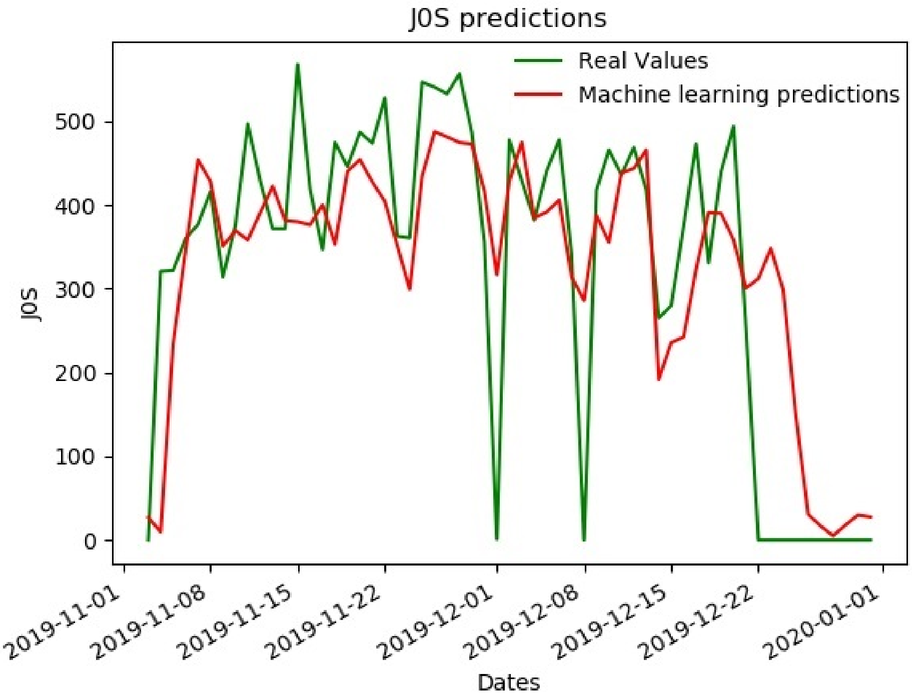


Figure 5.4: Predictions model II

implemented in our solution.

The model IV for predicting the 5TA915105B(J0V) production is very similar to the ones described before and presents similar results and consists on one LSTM layer with 2000 units and a dropout of 20% , we utilized a batch size of 64, and trained for 200 epochs. The model was compiled using the "Adam" optimizer and on figure 5.7 the training and validation errors (Y axis) evolution across epochs (X axis) are shown and on figure 5.8 the predictions of production of cars with the 5TA915105B batterie in the last 62 days of 2019 days in green as well as the true values in red are shown.

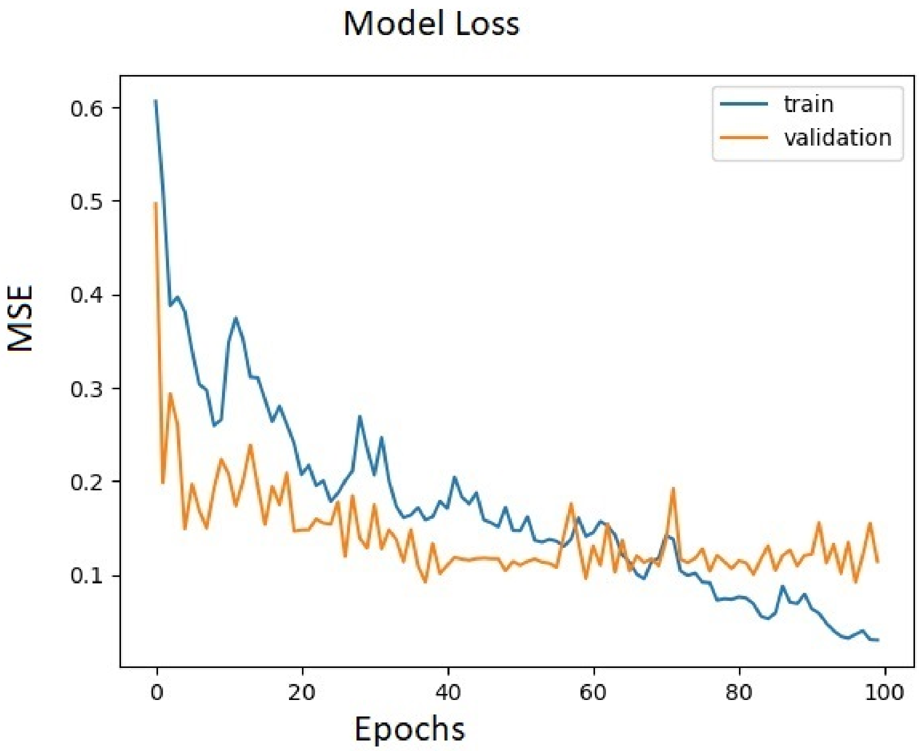


Figure 5.5: Error model III

This shows that the machine learning layer can predict the production with a although we have only 2 years of data to train the model and we expect the model's performance to improve significantly with the addition of more data and even though some different models and parameters were tested this is still an early stage of the developing phase and we expect improvements moving forward.

Now regarding the results from the entire system we can't exactly show actual results because the system needs to be put in place to be tested and then it would require multiple months until conclusions on the actual results can be analysed. The application of this system on historical data to evaluate its performance would be clearly biased so it was disregarded.

However historical data can be used to estimate the possible optimizations. In chapter 3 of this document we pointed out that in 2018 there were multiple cases of overstock of

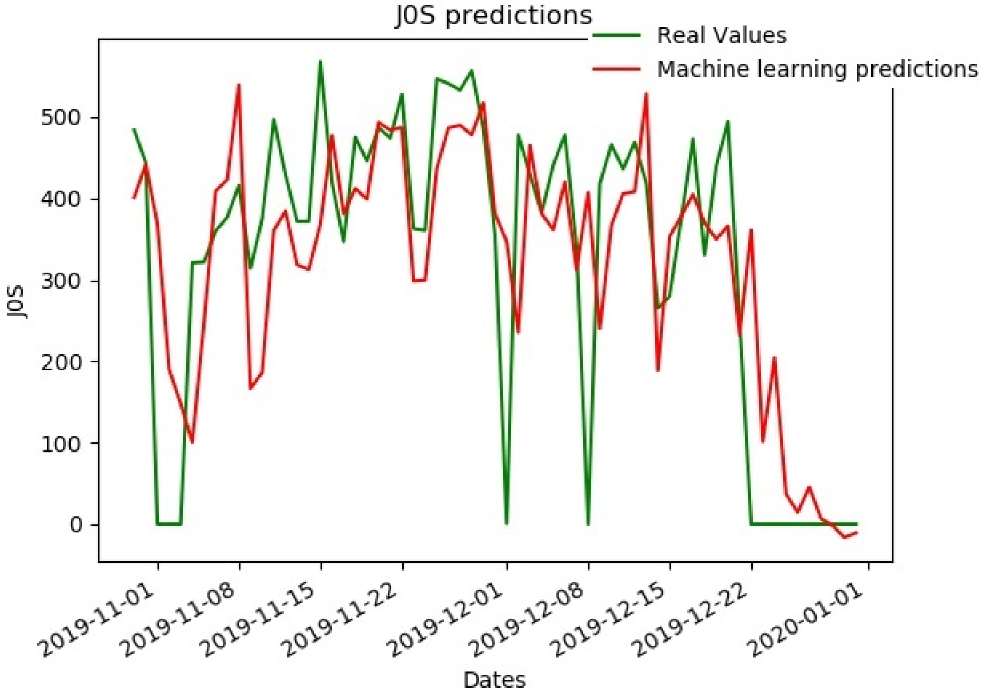


Figure 5.6: Predictions model III

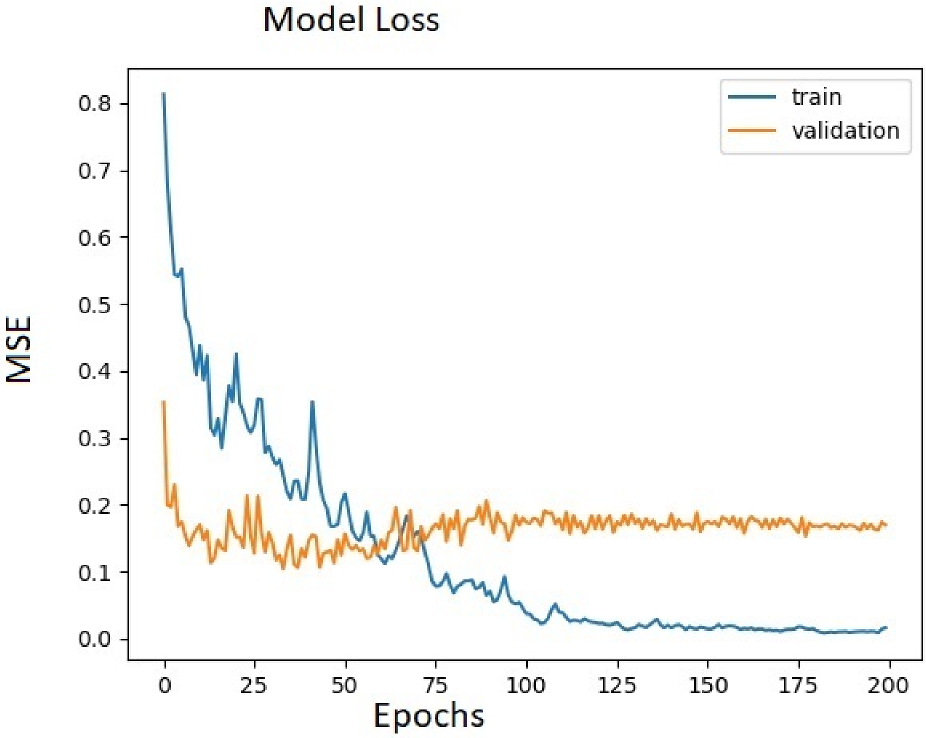


Figure 5.7: Error model IV

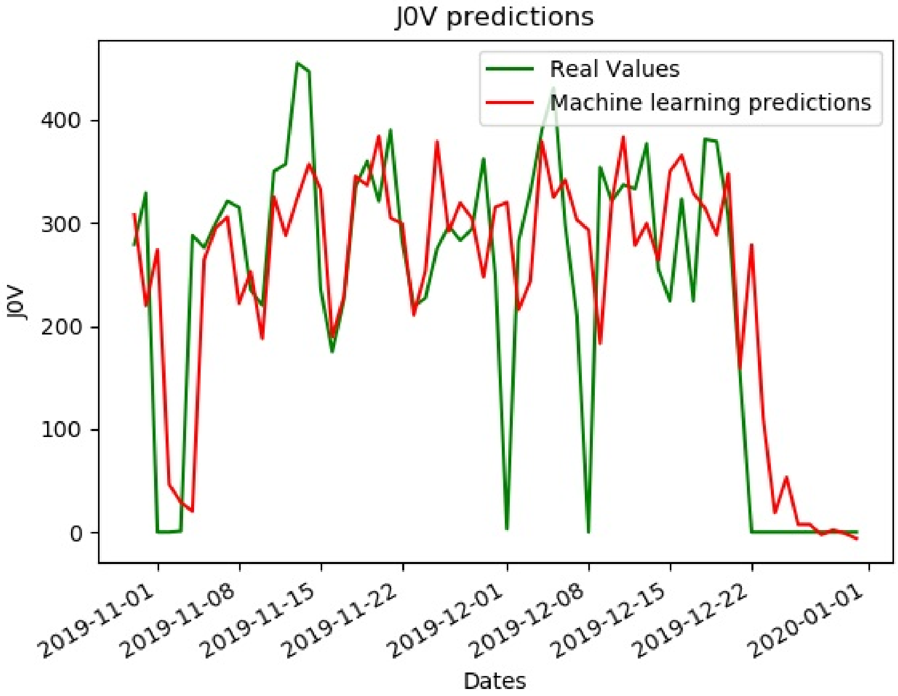


Figure 5.8: Predictions model IV

car batteries at VWAE, particularly in table 3.7 we can see that for every car batterie in VWAE we are at least half of the time in overstock situations and 25% of the time in severe overstock, to give some perspective the maximum inventory value of the 1S0915105A component in 2018 was 75 packages, if the factory was functioning at optimal pace (900 cars day) and only produced cars with the 1S0915105A batterie it would take more than 9 days to deplete the warehouse. If for example, we look at the situation in September illustrated in chapter 3 in figure 3.2 and by analysing this data in detail we can see how our system can reduce situations of overstock. The inventory levels for the 1S0915105A batterie starts the month at 26 packages and continues to drop to 21, that is still over the security level, until a truck arrives with 21 packages on day 2 increasing the level to 42, and again on the forth 26 more packages and again on the fifth 10 more. During these 4 days arrived 57 packages and were consumed around 25. This led to a value of 65 packages on the end of the fifth of September. Even after this peak the inventory levels drop for a few days but never dip below the overstock threshold and then a new mountain of overstock where the maximum value reached 61 packages. The orders of this material were placed 5 days before their arrival at VWAE, we believe that if planners get access to this data and our visualization capabilities as we can see in figure 5.9 and our machine learning predictions situations like the one described before would occur less frequently and with less impact. With transport, inventory and production data available as well as our machine learning predictions in a single platform with easy access we believe we can help logistics planners at VWAE reduce the occurrence of overstock situations in half regarding car batteries. This will save warehouse space and reduce warehouse costs.

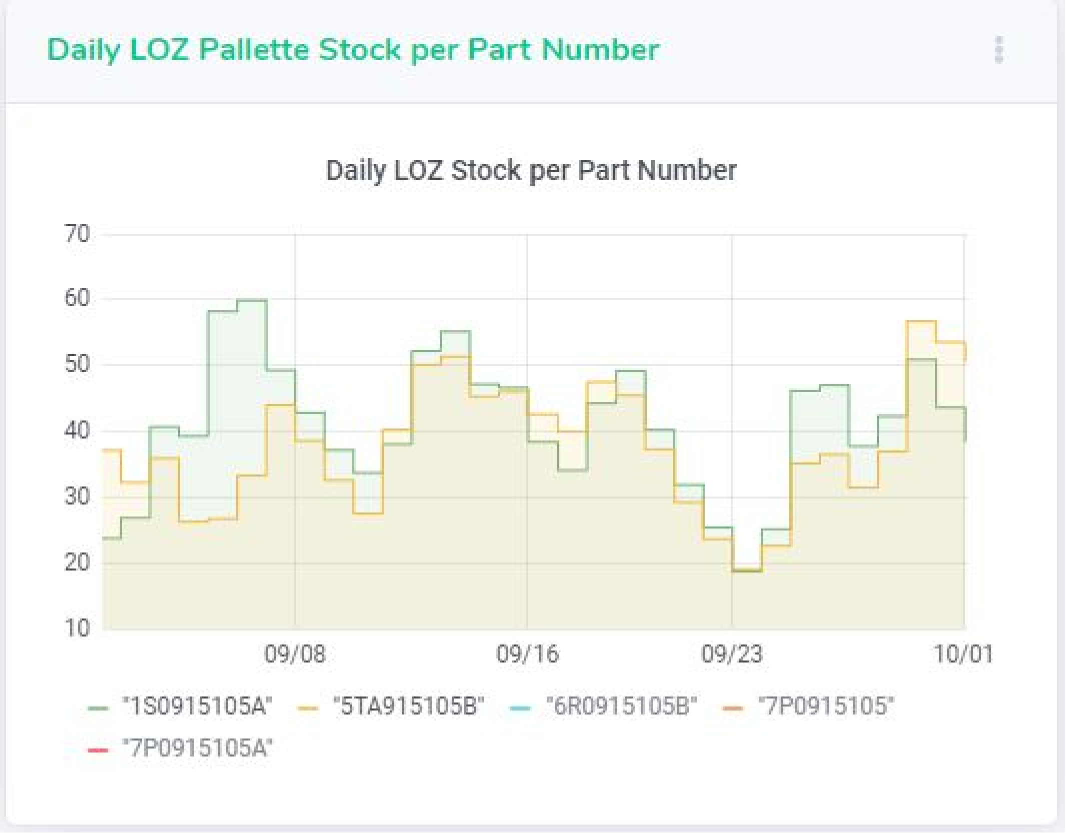


Figure 5.9: Inventory levels on the BOOST 4.0 Dashboard

Chapter 6

Conclusions and Future work

6.1 Conclusions

The main objectives of this work were to improve efficiency in intralogistics processes at VWAE and to present a data-driven system capable of reducing situations of overstock. We believe that both objectives were accomplished.

Regarding the machine learning model, we feel good about the choice of a LSTM model as it quickly noticed patterns in a time series data.

The decision to do an internship at VWAE during this project was in my opinion positive for both sides. I learned a lot during this internship and was able to apply some of the skills that I learned during my academic path. I also believe that the work I did at VWAE may prove to be an asset, especially the work focused on the data gathering automatization, and I believe that I contributed to a better data culture at VWAE and that in the future this will bring great advantages to VWAE. The toughest objective during this project was the integration of data between the multiple sources and the subsequent elimination of data silos and this challenge could have been even more difficult had it not been for the constant support that I enjoyed thanks to the internship. The fact that I can talk to people who really know the systems in place and that I can validate data on the shop floor has proven important.

Looking at the whole scope of the BOOST 4.0 project this system is only one way we can improve intralogistics processes and we see the actions in the data gathering and data integration as the biggest contribute made for improvement in the present and future. This integration process was made with the purpose of being scalable and this can be the groundwork for multiple future applications of data-driven solutions.

This now allows management at VWAE to select a focus problem and with minimal effort from planner's, powerful data-based solutions can be quickly built.

The contribute to the creation of a digital twin of the intralogistics processes at VWAE also added value to the final solution and opened paths for more innovative and interactive optimizations.
The possibility of improvements in the intralogistics process planning with data-driven systems is demonstrated and this system can be used as experience to build more solutions without as much initial resistance created by the necessity of data quality, quantity and integration.

6.2 Future work

As stated before, many alternative optimizations and systems can follow so as future work, there are multiple different paths: The first optimization is to our machine learning model because this is a data hungry model we could improve the results if we keep on increasing the data that we feed the model. This is not the only thing that could improve the results, the process of building a machine learning model is iterative and although our model is successful it is still not close to optimal. There are multiple things that can be adjusted, like the input data by adding features, the number of layers in the model, or parameters like batch sizes or the number of epochs to train the model.
Another path is the scalability of this implemented system for inventory optimization of car batteries to the entire scope of car components present in VWAE. The foundations for this work are already built since the entire system was built considering this possibility.
This system is built to look for optimizations regarding the functioning of the processes, but a financial component can be added to look for optimizations in a more lucrative way. As an example, with the addition of some data the system could be utilized to minimize the logistics costs(transport + handling+ storage costs) directly instead of the number of transport trucks or stock levels.
The data gathering and understanding showed that there is still immense room for optimizations since there is a lot of unexplored data throughout the different clusters on the VWAE factory.
Another possible future work is the search of patterns within the multiple clusters of data by comparing it with some available data like the weather conditions per example.
The preparing and feeding data process to the digital twin of the intralogistics processes studied is also a path that has multiple ramifications. The scalability to all car components seems the most natural first step.

Bibliography

[1] S. Vaidyaa. "Industry 4.0 and the current status as well as future prospects on logistics." In: *Procedia Manufacturing* 20 (2018), pp. 233–238. ISSN: 2351-9789.

[2] D. Reinsel, J. Gantz, and J. Rydning. "The Digitization of the World - From Edge to Core." In: *Framingham: International Data Corporation* November (2018), US44413318. URL: https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf.

[3] Q. QI and F. TAO. "Digital Twin and Big Data Towards Smart Manufacturing and Industry 4.0: 360 Degree Comparison." In: *IEEE Access* 6 (2018), pp. 23–34. ISSN: 3585-3593. DOI: http://www.ieee.org/publications_standards/publications/rights/index.html.

[4] P. Simon. *The visual organization: Data visualization, big data, and the quest for better decisions*. John Wiley & Sons, 2014.

[5] M. Holler, F. Uebernickel, and W. Brenner. "Digital Twin Concepts in Manufacturing Industries - A Literature Review and Avenues for Further Research." In: *Proceedings of the 18th International Conference on Industrial Engineering (IJIE)*, Korean Institute of Industrial Engineers: Korean Institute of Industrial Engineers, 2016. URL: https://www.alexandria.unisg.ch/249292/.

[6] T. Wuest, D. Weimer, C. Irgens, and K.-D. Thoben. "Machine learning in manufacturing: advantages, challenges, and applications." In: *Production & Manufacturing Research* 4.1 (2016), pp. 23–45. DOI: 10.1080/21693277.2016.1192517. eprint: https://doi.org/10.1080/21693277.2016.1192517. URL: https://doi.org/10.1080/21693277.2016.1192517.

[7] E. Hofmann and M. Rusch. "Industry 4.0 and the current status as well as future prospects on logistics." In: *Computers in Industry* 89 (2017), pp. 23–34. ISSN: 0166-3615. DOI: http://dx.doi.org/10.1016/j.compind.2017.04.002.

[8] N. Schmidtke, F. Behrendt, L. Thater, and S. Meixner. "Technical potentials and challenges within internal logistics 4.0." In: *2018 4th International Conference on Logistics Operations Management (GOL)*. 2018, pp. 1–10. DOI: 10.1109/GOL.2018.8378072.

[9] *D2.3 – Pilots description, adaptations and executive plans v1*. 2019. URL: https://cordis.europa.eu/project/id/780732/results.

[10] J. F. Krafcik. "Triumph of the lean production system." In: *MIT Sloan Management Review* 30.1 (1988), p. 41.

[11] J. P. Womack and D. T. Jones. "Lean thinking—banish waste and create wealth in your corporation." In: *Journal of the Operational Research Society* 48.11 (1997), pp. 1148–1148.

[12] N. Stefanovic and D. Stefanovic. "Supply chain business intelligence: technologies, issues and trends." In: *Artificial intelligence an international perspective*. Springer, 2009, pp. 217–245.

[13] A. Kadadi, R. Agrawal, C. Nyamful, and R. Atiq. "Challenges of data integration and interoperability in big data." In: *2014 IEEE International Conference on Big Data (Big Data)*. 2014, pp. 38–40.

[14] R. Wirth and J. Hipp. "CRISP-DM: Towards a standard process model for data mining." In: *Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining*. 2000, pp. 29–39.

[15] *D2.7 – Boost 4.0 standardization certification v1*. 2019. URL: https://cordis.europa.eu/project/id/780732/results.

[16] A. W. Colombo, T. Bangemann, S. Karnouskos, J. Delsing, P. Stluka, R. Harrison, F. Jammes, J. L. Lastra, et al. "Industrial cloud-based cyber-physical systems." In: *The Imc-aesop Approach* 22 (2014), pp. 4–5.

[17] D. Lukač. "The fourth ICT-based industrial revolution "Industry 4.0"— HMI and the case of CAE/CAD innovation with EPLAN P8." In: *2015 23rd Telecommunications Forum Telfor (TELFOR)*. 2015, pp. 835–838.

[18] G. Rathwell and P. Ing. "Design of enterprise architectures." In: *pera. net,[Online], Available: http://www. pera. net/Levels. html,[Accessed: Apr. 30, 2010]* (2004).

[19] P. B. Kruchten. "The 4+ 1 view model of architecture." In: *IEEE software* 12.6 (1995), pp. 42–50.

[20] M. Hankel and B. Rexroth. *Das Referenzarchitekturmodell Industrie 4.0 (RAMI 4.0)*. 2015.

[21] *DIN SPEC 91345:2016-04, Referenzarchitekturmodell Industrie 4.0 (RAMI4.0)*. DOI: 10.31030/2436156. URL: https://doi.org/10.31030/2436156.

[22] D. Knoll, M. Prüglmeier, and G. Reinhart. "Predicting Future Inbound Logistics Processes Using Machine Learning." In: *Procedia CIRP* 52 (2016). The Sixth International Conference on Changeable, Agile, Reconfigurable and Virtual Production (CARV2016), pp. 145 –150. ISSN: 2212-8271. DOI: https://doi.org/10.1016/j.procir.2016.07.078. URL: http://www.sciencedirect.com/science/article/pii/S2212827116308770.

[23] V. Gružauskas, E. Gimžauskienė, and V. Navickas. "Forecasting accuracy influence on logistics clusters activities: The case of the food industry." In: *Journal of Cleaner Production* 240 (2019), p. 118225. ISSN: 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2019.118225. URL: http://www.sciencedirect.com/science/article/pii/S0959652619330951.

[24] M. Takamatsu, N. Yamamoto, H. Kawachi, A. Chino, S. Saito, M. Ueno, Y. Ishikawa, Y. Takazawa, and K. Takeuchi. "Prediction of early colorectal cancer metastasis by machine learning using digital slide images." In: *Computer Methods and Programs in Biomedicine* 178 (2019), pp. 155 –161. ISSN: 0169-2607. DOI: https://doi.org/10.1016/j.cmpb.2019.06.022. URL: http://www.sciencedirect.com/science/article/pii/S016926071930197X.

[25] Lu and S. CY. "Machine learning approaches to knowledge synthesis and integration tasks for advanced engineering automation." In: *Computers in Industry* 15.1-2 (1990), pp. 105–120.

[26] S. B. Kotsiantis, I Zaharakis, and P Pintelas. "Supervised machine learning: A review of classification techniques." In: *Emerging artificial intelligence applications in computer engineering* 160 (2007), pp. 249–268.

[27] R. Cont and A. Kukanov. "Optimal order placement in limit order markets." In: *Quantitative Finance* 17.1 (2017), pp. 21–39. DOI: 10.1080/14697688.2016.1190030. eprint: https://doi.org/10.1080/14697688.2016.1190030. URL: https://doi.org/10.1080/14697688.2016.1190030.

[28] J. L. Berral, I. n. Goiri, R. Nou, F. Julià, J. Guitart, R. Gavaldà, and J. Torres. "Towards Energy-Aware Scheduling in Data Centers Using Machine Learning." In: *Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking*. e-Energy '10. Passau, Germany: Association for Computing Machinery, 2010, 215–224. ISBN: 9781450300421. DOI: 10.1145/1791314.1791349. URL: https://doi.org/10.1145/1791314.1791349.

[29] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich. "Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine." In: Omnipress. 2010.

[30] H. Sun, H. X. Liu, H. Xiao, R. R. He, and B. Ran. "Use of local linear regression model for short-term traffic forecasting." In: *Transportation Research Record* 1836.1 (2003), pp. 143–150.

[31] W.-I. Lee, C.-W. Chen, K.-H. Chen, T.-H. Chen, and C.-C. Liu. "Comparative study on the forecast of fresh food sales using logistic regression, moving average and BPNN methods." In: *Journal of Marine Science and Technology* 20.2 (2012), pp. 142–152.

[32] C. Sudheer, R Maheswaran, B. K. Panigrahi, and S. Mathur. "A hybrid SVM-PSO model for forecasting monthly streamflow." In: *Neural Computing and Applications* 24.6 (2014), pp. 1381–1389.

[33] C. Gao, E. Bompard, R. Napoli, and H. Cheng. "Price forecast in the competitive electricity market by support vector machine." In: *Physica A: Statistical Mechanics and its Applications* 382.1 (2007). Applications of Physics in Financial Analysis, pp. 98 –113. ISSN: 0378-4371. DOI: https://doi.org/10.1016/j.physa.2007.03.050. URL: http://www.sciencedirect.com/science/article/pii/S0378437107003251.

[34] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory." In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735. eprint: https://doi.org/10.1162/neco.1997.9.8.1735. URL: https://doi.org/10.1162/neco.1997.9.8.1735.

[35] F. A. Gers, N. N. Schraudolph, and J. Schmidhuber. "Learning precise timing with LSTM recurrent networks." In: *Journal of machine learning research* 3.Aug (2002), pp. 115–143.

[36] J. Schmidhuber. "Deep learning in neural networks: An overview." In: *Neural Networks* 61 (2015), pp. 85 –117. ISSN: 0893-6080. DOI: https://doi.org/10.1016/j.neunet.2014.09.003. URL: http://www.sciencedirect.com/science/article/pii/S0893608014002135.

[37] Y. Tsai, Y. Zeng, and Y. Chang. "Air Pollution Forecasting Using RNN with LSTM." In: *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*. 2018, pp. 1074–1079. DOI: 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00178.

[38] F. Chollet et al. *Keras*. https://keras.io. 2015.

[39] K. Chen, Y. Zhou, and F. Dai. "A LSTM-based method for stock returns prediction: A case study of China stock market." In: *2015 IEEE International Conference on Big Data (Big Data)*. 2015, pp. 2823–2824. DOI: 10.1109/BigData.2015.7364089.

[40] K. Choi, G. Fazekas, and M. Sandler. "Text-based LSTM networks for automatic music composition." In: *arXiv preprint arXiv:1604.05358* (2016).

[41] X. Wu, X. Zhu, G. Wu, and W. Ding. "Data mining with big data." In: *IEEE Transactions on Knowledge and Data Engineering* 26.1 (2014), pp. 97–107.

[42] P. Russom et al. "Big data analytics." In: *TDWI best practices report, fourth quarter* 19.4 (2011), pp. 1–34.

[43] Y. Zhang, S. Ma, H. Yang, J. Lv, and Y. Liu. "A big data driven analytical framework for energy-intensive manufacturing industries." In: *Journal of Cleaner Production* 197 (2018), pp. 57 –72. ISSN: 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2018.06.170. URL: http://www.sciencedirect.com/science/article/pii/S0959652618318201.

[44] J. Moyne and J. Iskandar. "Big data analytics for smart manufacturing: Case studies in semiconductor manufacturing." In: *Processes* 5.3 (2017), p. 39.

[45] Y. Kang, I. Park, J. Rhee, and Y. Lee. "MongoDB-Based Repository Design for IoT-Generated RFID/Sensor Big Data." In: *IEEE Sensors Journal* 16.2 (2016), pp. 485–497.

[46] Y. Zhang, S. Ren, Y. Liu, and S. Si. "A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products." In: *Journal of Cleaner Production* 142 (2017). Special Volume on Improving natural resource management and human health to ensure sustainable societal development based upon insights gained from working within 'Big Data Environments', pp. 626 –641. ISSN: 0959-6526. DOI: https://doi.org/10.1016/j.jclepro.2016.07.123. URL: http://www.sciencedirect.com/science/article/pii/S0959652616310198.

[47] K. Witkowski. "Internet of Things, Big Data, Industry 4.0 – Innovative Solutions in Logistics and Supply Chains Management." In: *Procedia Engineering* 182 (2017). 7th International Conference on Engineering, Project, and Production Management, pp. 763 –769. ISSN: 1877-7058. DOI: https://doi.org/10.1016/j.proeng.2017.03.197. URL: http://www.sciencedirect.com/science/article/pii/S1877705817313346.

[48] D. Maltby. "Big data analytics." In: *74th Annual Meeting of the Association for Information Science and Technology (ASIST)*. 2011, pp. 1–6.

[49] S. Srinivasa and S. Mehta. *Big Data Analytics: Third International Conference, BDA 2014, New Delhi, India, December 20-23, 2014. Proceedings*. Vol. 8883. Springer, 2014.

[50] J. Zakir, T. Seymour, and K. Berg. "BIG DATA ANALYTICS." In: *Issues in Information Systems* 16.2 (2015).

[51] A. G. Shoro and T. R. Soomro. "Big data analysis: Apache spark perspective." In: *Global Journal of Computer Science and Technology* (2015).

[52] *MLlib: Apache Spark*. URL: https://spark.apache.org/mllib/.

[53] *The MongoDB 4.2 Manual*. URL: https://docs.mongodb.com/manual/.

[54] URL: http://cassandra.apache.org/doc/latest/.

[55] D. Sheakh. "A Study of Inventory Management System Case Study." In: *Journal of Dynamical and Control Systems* 10 (May 2018), pp. 1176–1190.

[56] J.-S. Song, G.-J. van Houtum, and J. A. Van Mieghem. "Capacity and Inventory Management: Review, Trends, and Projections." In: *Manufacturing Service Operations Management* 22.1 (2020), pp. 36–46. DOI: 10.1287/msom.2019.0798. eprint: https://doi.org/10.1287/msom.2019.0798. URL: https://doi.org/10.1287/msom.2019.0798.

[57] A. Dolgui and C. Prodhon. "Supply planning under uncertainties in MRP environments: A state of the art." In: *Annual Reviews in Control* 31 (Dec. 2007), pp. 269–279. DOI: 10.1016/j.arcontrol.2007.02.007.

[58] S. Koh and S. Saad. "MRP-controlled manufacturing environment disturbed by uncertainty." In: *Robotics and Computer-Integrated Manufacturing* 19 (Feb. 2003), pp. 157–171. DOI: 10.1016/S0736-5845(02)00073-X.

[59] S. Axsäter. "A Capacity Constrained Production-Inventory System with Stochastic Demand and Production Times." In: *International Journal of Production Research - INT J PROD RES* 48 (Oct. 2010), pp. 6203–6209. DOI: 10.1080/00207540903283808.

[60] J. Gijsbrechts, R. Boute, D. Zhang, and J. Van Mieghem. "Can Deep Reinforcement Learning Improve Inventory Management? Performance on Dual Sourcing, Lost Sales and Multi-Echelon Problems." In: *SSRN Electronic Journal* (July 2019). DOI: 10.2139/ssrn.3302881.

[61] N. Stefanovic, D. Stefanovic, and B. Radenkovic. "Application of Data Mining for Supply Chain Inventory Forecasting." In: *Applications and Innovations in Intelligent Systems XV*. Ed. by R. Ellis, T. Allen, and M. Petridis. London: Springer London, 2008, pp. 175–188. ISBN: 978-1-84800-086-5.

[62] N. Xue, I. Triguero, G. P. Figueredo, and D. Landa-Silva. "Evolving Deep CNN-LSTMs for Inventory Time Series Prediction." In: *2019 IEEE Congress on Evolutionary Computation (CEC)*. 2019, pp. 1517–1524.

[63] B. Hussein, A. Kasem, S. Omar, and n. z. Siau. "A Data Mining Approach for Inventory Forecasting: A Case Study of a Medical Store: Proceedings of the Computational Intelligence in Information Systems Conference (CIIS 2018)." In: Jan. 2019, pp. 178–188. ISBN: 978-3-030-03301-9. DOI: 10.1007/978-3-030-03302-6_16.

[64] S. Li and X. Kuo. "The inventory management system for automobile spare parts in a central warehouse." In: *Expert Systems with Applications* 34.2 (2008), pp. 1144 –1153. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2006.12.003. URL: http://www.sciencedirect.com/science/article/pii/S0957417406004015.

[65] T. pandas development team. *pandas-dev/pandas: Pandas*. Version latest. Feb. 2020. DOI: 10.5281/zenodo.3509134. URL: https://doi.org/10.5281/zenodo.3509134.

[66] Wes McKinney. “Data Structures for Statistical Computing in Python.” In: *Proceedings of the 9th Python in Science Conference*. Ed. by Stéfan van der Walt and Jarrod Millman. 2010, pp. 56 –61. DOI: 10.25080/Majora-92bf1922-00a.

[67] 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, and E. Duchesnay. “Scikit-learn: Machine Learning in Python.” In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[68] *PostgreSQL: The world’s most advanced open source database*. URL: https://www.postgresql.org/ (visited on 06/06/2020).

[69] W. Ali, M. U. Shafique, M. A. Majeed, and A. Raza. “Comparison between SQL and NoSQL databases and their relationship with big data analytics.” In: *Asian Journal of Research in Computer Science* (2019), pp. 1–10.

[70] T. N. Khasawneh, M. H. AL-Sahlee, and A. A. Safia. “SQL, NewSQL, and NOSQL Databases: A Comparative Survey.” In: *2020 11th International Conference on Information and Communication Systems (ICICS)*. 2020, pp. 013–021.

[71] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. “Dropout: a simple way to prevent neural networks from overfitting.” In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.