Tampere University of Applied Sciences

Development of a Warehouse Utilization Data Analysis and Forecast Tool

for Volkswagen AG

Christopher Specht

BACHELOR'S THESIS
May 2021

Degree Programme in Media and Arts
Interactive Media

ABSTRACT

Tampereen ammattikorkeakoulu
Tampere University of Applied Sciences
Degree Programme in Media and Arts
Interactive Media

SPECHT, CHRISTOPHER:
Development of a Warehouse Utilization Data Analysis and Forecast Tool

Bachelor's thesis 84 pages, of which appendices 7 pages
May 2021

Against the background of the fourth industrial revolution, Industry 4.0, production systems and warehouses increasingly get networked. These cyber-physical systems generate an exponentially rising amount of data. Consequently, the possibilities for deriving information for decision-making from the so-called big data increase as well.

The objective of this bachelor's thesis was to exploit one of these potentials for the automobile manufacturer Volkswagen AG. Logistics data from the Volkswagen Industrial Cloud was analyzed in order to obtain information about the utilization and stock development of multiple Volkswagen warehouses.

The required data was accessed and identified within several corresponding databases, following which datasets were created. A cloud computing-based calculation algorithm was developed afterward for evaluating the created datasets. Due to the complexity of the information, a visualization concept according to lean user experience design principles was created. This concept aimed to improve the efficiency of information comprehension for a human user.

For processing the above-mentioned project stages, the agile project management method Scrum was applied. The information needed was derived according to the intelligence cycle within the framework of structured data analysis.

The project work led to 100% data accuracy with regard to the stored stock of a sample warehouse in Hanover while other warehouses could be analyzed successfully afterward. However, the utilization accuracy still was to be monitored and tested at the time of project completion.

Furthermore, premises for future forecasting through AI-based software as a service tools were created. The data basis for this implementation needs to be collected over time as a subsequent step.

This thesis contains confidential information from Volkswagen AG. The affected sections were therefore blackened in the public report.

Key words: big data analysis, cloud computing, internet of things, logistics, industry 4.0

AUTHOR'S STATEMENT

Hereby I, Christopher Specht, assure that I have prepared the present bachelor's thesis independently, have used no aids other than those indicated and have marked the passages of the thesis that were taken in the wording or essential content from other works with exact indication of the source.

Tampere, 17.05.2021 C. Specht

BLOCKING NOTE

This bachelor's thesis is written in cooperation with the Volkswagen AG and is confidential. Therefore, publications and reproductions, even partial, are not permitted without the express permission of the Volkswagen AG. The access to the unblackened version shall only be made available exclusively to official thesis supervisors.

CONTENTS

ABBREVIATIONS AND TERMS

κ	utilization
AG	Ger.: Aktiengesellschaft, Eng.: stock corporation
AMS	automatic miniload system
AWS	Amazon Web Services
BI	business intelligence
BO	(SAP) Business Objects
CA	certification authority
CPS	cyber-physical system
CRN	container reference number; data field
CSS	cascading style sheet
CSV	comma separated values
d	depth; parameter
deldate	deletion date; data field
e	storage place, storage element; parameter
EDA	exploratory data analysis
h	height; parameter
HCI	human-computer interaction
HTML	hypertext markup language
IaaS	infrastructure as a service
IoT	Internet of Things
IT	information technology
JIS	just-in-sequence
JIT	just-in-time
KPI	key performance indicator
l	length; parameter
LDL	(Volkswagen) Logistics Data Lake
[redacted]	[redacted]
[redacted]	[redacted]
M2M	machine-to-machine
mflstat	material flow status; data field
n	number, amount, capacity; parameter
ODS	operational data store

OEM	original equipment manufacturer
PaaS	platform as a service
PBP	product backlog point
PDC	purchasing and distribution center
PDWHID	package data warehouse ID; data field
PHP	hypertext preprocessor
PKI	public key infrastructure
PlntN	plant number; data field
PoLA	principle of least astonishment
PRN	package reference number; data field
PrtN	part number; data field
RA	registration authority
RFID	radio frequency ID
s	stacking factor; parameter
S3	(Amazon) Simple Storage Service
SA	storage area; data field
SaaS	software as a service
SL	storage layer
SpplN	supplier number; data field
SQL	structured query language
st	stacks; parameter
SVG	scalable vector graphics
t	container type; parameter
UID	usage indicator; data field
update	update date; data field
UX	user experience
V	volume; parameter
VW CV	Volkswagen Commercial Vehicles
w	width; parameter
x	certain container dimension, either width or depth; parameter
XLSX	Excel spreadsheet XML
XML	extensible markup language

LIST OF FIGURES

LIST OF TABLES

LIST OF FORMULAS

1 INTRODUCTION

1.1 Topic relevance and current situation

Especially in the industrial sector, numerous optimizations and productivity increases are realized based on newly welling up technologies. These technologies include terms like big data, cloud computing, and the Internet of Things. Smart factories are supplied just-in-time to save storage space and capital commitment costs. Most individual customer wishes can be implemented without any significant delay by sharing and processing order data with suppliers in real-time via cloud services. Suppliers communicate with each other online to coordinate who must deliver which components where and when. Moreover, logistics centers bundle material flow streams from various suppliers to provide resources required for just-in-time production in line with demand.

As one of the world's largest automobile manufacturers and industrial companies, Volkswagen AG is naturally dedicated to topics of the so-called fourth industrial revolution (Industry 4.0). Networking in the context of digitalization is creating much more data exponentially. This amount of data could only be stored and processed by a company's own server system cost-intensively. An established method for managing this big data is renting cloud computing systems in conjunction with platform as a service (PaaS) or software as a service (SaaS) from external, specialized providers. The market development for cloud computing services in Germany reflects this trend. In comparison to 2011, for 2021 the market for these services is expected to be increased by factor 16 with an exponential tendency (Statista 2021b). Accordingly, there are even more opportunities for evaluating and analyzing these extraordinarily detailed datasets.

1.2 Thesis aim, methods, and structure

One of these potentials for information derivation is exploited by this bachelor's project: A utilization data analysis and forecast for Volkswagen AG logistics centers using the practical example of the purchasing and distribution center (PDC)

Hanover North of Volkswagen Commercial Vehicles (VW CV). This information about utilization is of enormous relevance for ensuring supply security for the company's production. Operational (short-term) questions that this information should answer are primarily: Is the stock of type X components sufficient for the following hours of just-in-time production? Is the warehouse over-utilized or under-utilized, thereby generating avoidable costs? And how will the stock utilization change by tomorrow, next week, and next month? Consecutive strategic (long-term) questions follow. Is the warehouse space planned for the warehouse sufficient? Can warehouse space be reduced in the long term, or may it need to be expanded? And are there identifiable container relocation processes that would improve utilization?

The aim of this bachelor's thesis is to generate information from the logistics data of Volkswagen AG in order to be able to answer the operational questions listed above as precisely as possible. In addition, the prerequisites for collecting data are to be created from which the information necessary for answering the strategic questions can be derived in the future.

In order to achieve this goal, methods from the field of big data analysis are applied primarily. In addition, due to the high level of complexity of the results, the abstraction of information is crucial for a human user. Against the background of decision efficiency, this thesis consequently also deals with the information design of the insights gained.

Theoretical backgrounds from the areas of logistics as well as data management and accessibility are required to achieve the mentioned targets. These two areas are meaningfully linked in the context of the already mentioned Industry 4.0. Therefore, this thesis's first subchapter (2.1) deals with the Industry 4.0 technologies needed for this project, which relate to logistics and data management. A section on trends in logistics closely interwoven with data management follows (Chapter 2.2). Since the central element of this project is the data of a purchasing and distribution center, the logistics center itself and its contributors, i.e., suppliers and manufacturers, are also put into context. All issues that substantially impact the type of data from the warehouse, including the supplier structure and delivery types, are explained in this subchapter. The following section (Chapter

2.3) is all about digital data. First, databases, their essential functions, and terminology are explained. Since a large part of the project revolves around programming a section on software development follows. After that, the visual-interactive preparation of digital data and information derived from it is explained. Lean user experience design offers corresponding methods for that. The theoretical part of the thesis concludes with a chapter (2.4) about the applied methodologies. It starts with the structured data analysis in conjunction with its intelligence cycle on which the algorithm development for the calculation of the warehouse utilization is based. Last, the agile project management method Scrum is described.

The second part of the thesis deals with the practical implementation of the theoretical approaches discussed beforehand. First (Chapter 3.1), the initial situation is presented based on the digitization strategies of Volkswagen AG. Important topics here are the classification of the project within Volkswagen logistics, the Amazon Web Services (cloud computing provider) tools used, and the data generation in the purchasing and distribution center. The chapter concludes with a detailed problem and use case description. The development of the analysis tool (chapter 3.2) is divided into six consecutive steps. The project initialization (3.2.1) and the requirement analysis (3.2.2) form the preliminary work for the execution of the four following so-called development sprints of Scrum management. These are namely data identification (3.2.3), dataset creation (3.2.4), data transfer and algorithm development (3.2.5), and finally, information design (3.2.6). The practical part concludes with Chapter 3.3, which focuses on preparing the utilization forecast and an outlook on the use of even more far-reaching big data technologies. The thesis is completed with a conclusion that, on the one hand, deals with the critical reflection and evaluation of the achieved results (Chapter 4.1). On the other hand, it summarizes the project and the contents of this thesis (Chapter 4.2). A complete visual overview of all in this thesis mentioned topics can be found in Appendix 1.

2 THEORETICAL BACKGROUND

2.1 Industry 4.0 as a leading trend in the automobile industry

The term "Industrie 4.0", or Industry 4.0 in English, has its origin in the 'research union economy-science' of the German federal government, which operated from 2006 to 2013. As part of this initiative, which was then called a future project, German industrial production (mainly containing automobile production) was to be linked with modern information and communication technology. The aim should be an utterly self-organized production (smart factories) that can be achieved through extensive networking. These smart factories should optimize an entire value chain instead of just a single production step. Therefore, the union announced recommended courses of action regarding the fourth industrial revolution on the Hannover Messe (an international industry fair) in 2013. (Platform Industrie 4.0 2021.)

A graphic representation of the four industrial revolutions is shown in Figure 1. While the first two industrial revolutions are completed in most parts of the world, the third industrial (or digital) revolution is still going on. It is primarily characterized by the widespread use of computers and automation of production systems. However, it was not until the fourth industrial revolution that the focus was on so-called cyber-physical systems (CPS), the Internet of Things (IoT), and overall, networking. (Ximea n.d.) A graphic overview of the technologies associated with Industry 4.0 is shown in Figure 2. All these technologies can also be found in the four design principles of Industry 4.0 defined by Hermann, Pentek & Otto (2016). Namely, these are interconnection, information transparency, decentralized decisions, and technical assistance. (Hermann et al. 2016.)

The principle of interconnection, for which the IoT and CPS form the basis, is of particular importance for this project. A basis on which machines, devices, sensors, and people can communicate in real-time must be created. Common communication standards are required to enable a flexible combination of machines from different suppliers. Such flexibility through the combinability of modules creates the necessary prerequisites for establishing smart factories of Industry 4.0.

These can react almost in real-time to the constantly changing requirements of the market. (Hermann et al. 2016.)

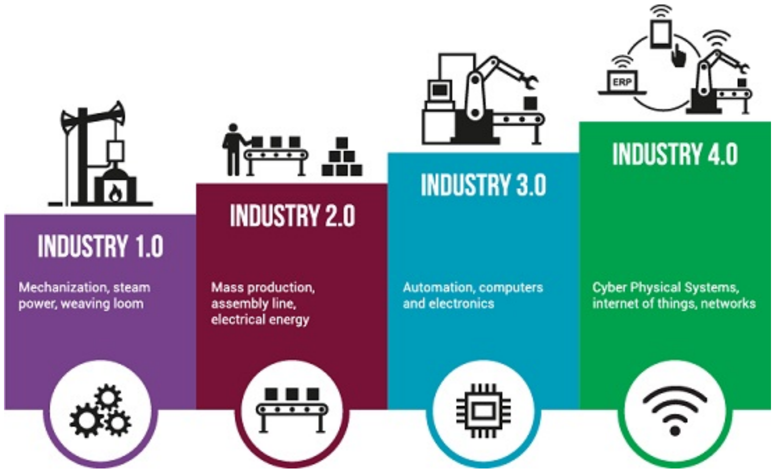

ERP
INDUSTRY 1.0
Mechanization, steam power, weaving loom
INDUSTRY 2.0
Mass production, assembly line, electrical energy
INDUSTRY 3.0
Automation, computers and electronics
INDUSTRY 4.0
Cyber Physical Systems, internet of things, networks

FIGURE 1. The four industrial revolutions (Ximea n.d.)

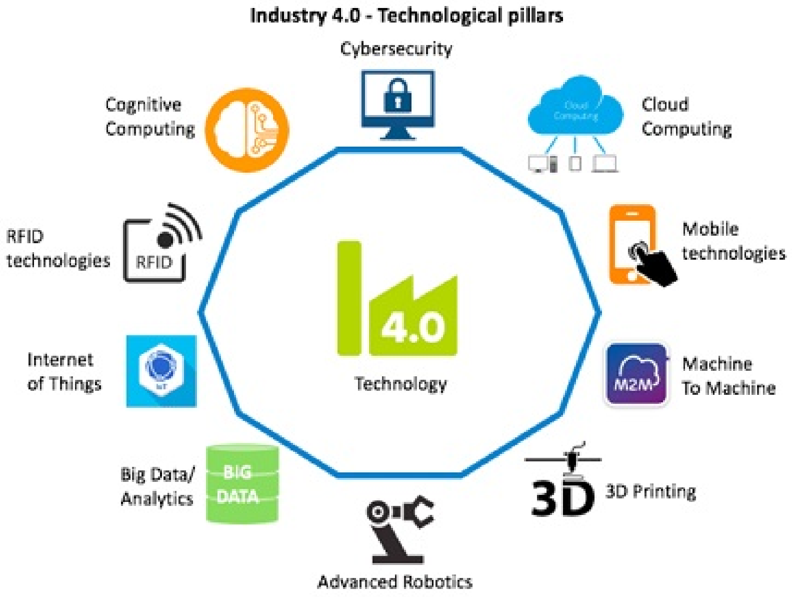


FIGURE 2. Industry 4.0 associated technologies (Ximea n.d.)

2.1.1 Internet of Things and cyber-physical systems

The terms Internet of Things (IoT) and cyber-physical system (CPS) often go hand in hand in the industrial environment, especially in the automobile industry. They describe the two sides of innovative, object-based networking technology. The CPS concentrates more on the description of the objects in the system themselves. They are mechanical, mainly independently operating components connected to one another via matching networks (the IoT). The aim is to be able to control, regulate and manage complex infrastructures. A central component of this technology is the real-time exchange of information between the networked objects. However, it does not matter whether this information exchange is wired or wireless in the true sense of the definition. (Luber 2017.)

CPSs are currently already being used to implement intelligent power grids, intelligent production systems, and in medical technology. One of the most consequential advantages of CPSs and the IoT is that the components, such as individual heavy-duty industrial robots in an entire production facility, can coordinate their actions with one another, i.e., have comprehensive communication capabilities. Significant prerequisites for implementing a CPS are sensors and actuators on or in the components themselves as well as a suitable network infrastructure with real-time or near-real-time processing (for example 5G networks). In conjunction, databases for the intermediate storage of, e.g., relevant production data, are needed. In most use cases, the components are therefore integrated into a cloud architecture and can therefore also access external computer resources such as additional processors or storage capacities. (Luber 2017.)

The IoT, on the other hand, describes the networking of the components. The components, also known as smart devices, are assigned unique identities (addresses) via which they can later be controlled from any other component or access point to the IoT (often via a browser-based graphical user interface). Another term that is often used for this context is machine-to-machine communication, or M2M for short. Radio frequency identification (RFID) is still a frequently used method for identifying and communicating individual objects. With this technology, so-called transponders are attached to an object that can be wirelessly recognized and read by networked, possibly even mobile readers at various production or warehouse points. The recorded data is then forwarded to a linked database in the IoT, from where it can be processed and retrieved again. (Luber 2016.)

2.1.2 Big data and big data analytics

The term big data describes larger and more complex datasets that accumulate at high speed. That is why it is also spoken of the three Vs of big data. These are variety, volume, and velocity. Conventional data processing software or classic spreadsheet programs can no longer process these extensive amounts of data. Because these datasets are that diverse, they hold enormous potential for analyses and evaluations to optimize business processes. Therefore, the demand for big data analytics systems has arisen. (Oracle n.d.b.)

As described in more detail in Chapter 2.4.1, the main goal of statistical evaluation of large amounts of data is to generate a better basis for decision-making for operational as well as strategic corporate planning. To this end, the data scientists, i.e., experts in big data analytics, take care of the integration, management, and evaluation of the data as part of the knowledge process. Computer-generated graphics or abstract depictions often help humans to derive knowledge or cause-and-effect principles from big data. In terms of administration, rented cloud systems, in which data is stored decentralized, help to avoid disproportionately high investment costs for in-house server extensions. Logically and contextually connected extensive collections of data within big data are also called data lakes.

New technology for the speed-optimized evaluation of the data is a hardware optimization called in-memory computing. Advances in storage technology and more powerful processors make it possible to load larger amounts of data directly into the computer's main memory executing the evaluation algorithm. That eliminates a decisive slowing factor, namely, the connection to slower storage media such as hard drives or databases. (Radtke 2019; Oracle n.d.b.)

2.1.3 Cloud computing

In a general sense, the demand-based provision of IT resources via the internet is understood as cloud computing. A unique feature here is that only usage-dependent prices are incurred, e.g., a fixed amount for the commissioning of the search of one terabyte of data. As already mentioned in the previous section, the decisive advantage is that there is no need to maintain own huge computing and server centers. Flexible access to existing cloud architecture is possible, which is usually set up by specialized companies. Computing power and storage capacity can be adapted flexibly to own requirements (advantage of scalability). (Amazon Web Services n.d.e.)

A distinction is made between three types of cloud computing. The first stage is infrastructure as a service (IaaS). Access to essential external network functions, computers, and data storage is granted. Because there are hardly any predefined links or calculation methods, IaaS offers the maximum degree of customization options and flexibility since mainly the hardware resources and their connection are made available. The next stage is the platform as a service (PaaS), which already contains operating systems for managing the infrastructure made available through IaaS. The service provider proactively takes over tasks such as resource procurement, data capacity planning, software maintenance, and patching (updates for the corresponding software). Software as a service (SaaS) is the perfection of cloud computing. Presumably, SaaS also is the variant with which private end users come into contact most because it includes complete products from the service provider, such as web-based e-mail programs. As part of SaaS, there is also no need for in-house software-on-software development so that only the operation of the corresponding product is in the foreground. (Amazon Web Services n.d.e.)

2.2 Trends in logistics and supply chain management

Current developments in logistics are shaped by headlines such as artificial intelligence, machine learning, and 5G networks. Cyber-physical systems are networked in real-time, generate large amounts of data, and act mostly autonomously. Secure data rooms are necessary to be able to carry out big data networking in a protected atmosphere. This broad range of new technologies is also called the silicon economy, the basis for an industrial platform economy. (Vogel-Heuser, Bauernhansl & ten Hompel 2020, 3.)

For the first time in the history of industrial development, computer performance is not necessarily the limiting factor in exploiting new optimization potential. Rather, considerations have to be made as to how the available resources are to be ideally used. Innovative algorithms are to be constructed to connect the growing number of distributed systems of the future Internet of Things and use them to the fullest.

Logistics in particular is one of the leading application sectors of these new technologies because the basic model of logistics, i.e., bringing the correct goods to the right place at the right time, is described as entirely deterministic and algorithmizable. Another advantage is that the basic building blocks of logistics can be characterized thoroughly and are accordingly often standardized. (Vogel-Heuser et al. 2020, 4.) However, that does not necessarily mean the information derivation from logistics data is simple, as ten Hompel states.

> The complexity arises as a result of many process steps and their simultaneous, flexible networking and, above all, their multi-criteria optimization. The disposition of logistic distribution systems is one of the most complex tasks of modern IT, and the high-frequency internet trade has taken care of the rest. (Vogel-Heuser et al. 2020, 5.)

Often supply chain management is mentioned as an essential trend in logistics, too. In its original meaning, not only the company's own value chain is considered but the entire value chain, i.e., from the source of supply to the point of consumption (Stephens 1989, 3–8). Just-in-time and just-in-sequence (Chapter 2.2.2) production are used, and inter-organizational rationalization potentials are exploited. An important step in that direction is coupling the respective production and logistic information systems of suppliers and manufacturers. (Weber 2012, 19–23.)

2.2.1 The suppliers' pyramid in the automobile industry

Vehicle manufacturers (e.g., BMW, Daimler, Volkswagen) in the automotive industry are – often against the background of logistics processes – referred to as original equipment manufacturer, OEM for short. OEMs sell their products either via direct but mostly via indirect (mainly car dealerships) distribution channels under their own brand names to end consumers. However, the vehicles also largely contain products from other manufacturers of modules, components, or individual parts (e.g., Bosch, Continental, ZF). (Hundertmark 2013, 1.)

Depending on the degree of complexity and proximity to the end product or OEM, the suppliers are categorized into three levels of a so-called supplier pyramid, as Figure 3 shows.

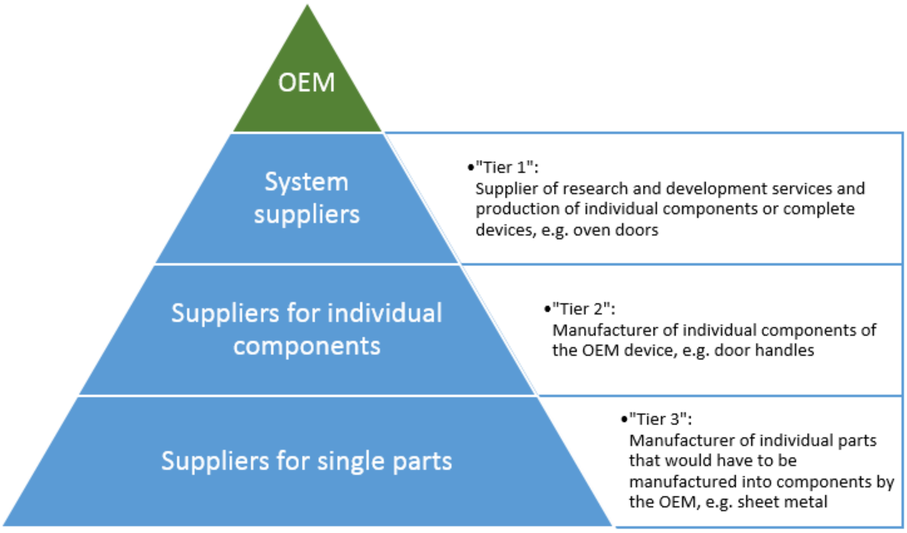

OEM
System suppliers
Suppliers for individual components
Suppliers for single parts
•"Tier 1":
Supplier of research and development services and production of individual components or complete devices, e.g. oven doors
•"Tier 2":
Manufacturer of individual components of the OEM device, e.g. door handles
•"Tier 3":
Manufacturer of individual parts that would have to be manufactured into components by the OEM, e.g. sheet metal

FIGURE 3. The suppliers' pyramid and its tiers (Gronbach 2018)

The suppliers for individual parts, generally called third-tier suppliers, are at the lowest level and thus also the broadest. They manufacture parts such as sheet metal or screws that have to be made into components by the OEM. They are followed by the suppliers for individual components (second-tier suppliers) who can already deliver individual components of an end product, e.g., door handles. The system or module suppliers (first-tier suppliers) are closest to the OEM. In some cases, they can independently take on research and development services, and produce systems and modules that already operate independently, such as brake systems, distance warning systems, or tire sets. (Hundertmark 2013, 4.)

The trend is that OEMs are increasingly turning to first-tier suppliers to assign them far-reaching development tasks. That enables a shortening of the time to market for new vehicles and often also means more cost-effective production due to the competition between the suppliers. However, there are also obstacles and novel costs as described in the following section. (Gronbach 2018.)

2.2.2 Delivery variations for production processes

The only decisive factor for the delivery form for production is the original equipment manufacturer, who defines and specifies the corresponding concepts via the supplier contracts. The logistics concepts were particularly shaped and determined by the automotive industry. The rapidly growing number of variants and the associated diversification of customer orders made it necessary to develop the delivery concepts further. Above all, a greater variety of variants led to an increasing divergence of the modules, components, and individual parts. Therefore, it cannot be stuck to a classic ordering method. It would cause massive expansion of storage and logistics space, personnel, and operating resources and thus a disproportionate increase in capital commitment and storage costs. (Hummel 2019, 4–5.)

Figure 4 shows a comparison of the four delivery variants established today in the automotive industry, which can mainly be distinguished in the number of suppliers and the proportion of the total delivery volume per supplier. Further comparison criteria, derived from the two core targets of logistics, are the required average space and time requirements. (Weber 2012, 4.)

The high number of suppliers for conventional warehouse delivery offers both advantages and disadvantages. In general, there is less dependency on a single supplier since, in the event of a delivery failure, a fallback to an alternative third-tier supplier is likely possible. In addition, due to the high level of competition between suppliers, a lower average purchase price per component can be expected. However, this flexibility has its downside. Often there are no long-term contractual arrangements between OEM and supplier. Consequently, offers must be requested, checked, and assessed regularly. Furthermore, delivery times must be coordinated more frequently.

Although higher minimum stocks allow more flexible appointments, as the complexity of the parts ordered increases, this can also result in a costly expansion of storage capacities. In addition, the advantage of less dependency due to more specific components is also becoming less important. That is because there are usually only a few and, in extreme cases, only a single supplier capable of delivering specific module components, such as vehicle door locks.

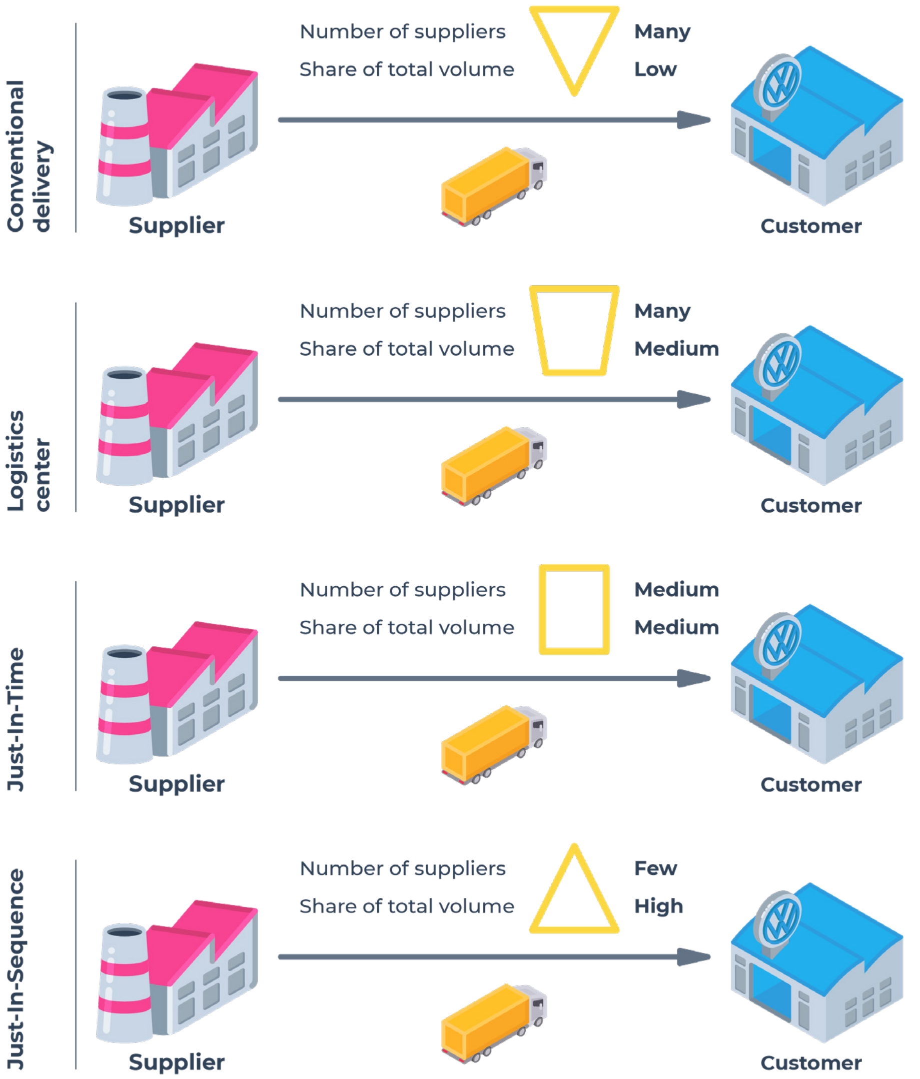


FIGURE 4. Delivery variations in logistics (Hummel 2019, 3, modified)

External logistics centers have established themselves, particularly in order to level out existing time disparities. Additional service providers often act as an interface between the OEM and the supplier. They manage the stocks and, ideally, guarantee a 24-hour delivery option. The OEM reports monthly, weekly, daily, and hourly requirements whereupon appropriate transports are organized. This system has the potential to equalize conventional ordering rhythms significantly. With a corresponding contractual arrangement, the responsibility for own on-time and quantity-based production delivery can even be assigned to the service provider. It can also make sense to position a logistics center centrally between several supplier locations in order to bundle material and goods flows and thus reduce transport costs. (Hummel 2019, 4–5.)

The next level of delivery concepts is just-in-time delivery and production. With a balanced supplier-total volume ratio, the just-in-time delivery is coordinated explicitly in such a way that there is no need for large-scale warehouses for preliminary products and individual parts. Only relatively small intermediate buffers are maintained for any irregularities in the production flow. Since there are no capital tie-up costs (assuming sufficient production capacity and constant financial resources), the throughput and thus the productivity of production can be increased on the procurement side.

The currently most challenging delivery variant in terms of coordination is just-in-sequence delivery and production. Although only a few first-tier suppliers are actively involved in the production, they, conversely, have a decisive share of the total delivery volume. Since only highly specific and mostly pre-assembled modules are delivered, the OEM, in this case, is maximally dependent on the supplier. A delivery failure or strike usually leads to far-reaching consequences. For the OEM, this usually means that production cannot be continued. It may also be common for the OEM to transfer increased quality standards to second- and third-tier suppliers. For this reason, the manufacturer sometimes establishes the links between individual part suppliers and first-tier suppliers after negotiations with both. In the case of very large OEMs, it is not unusual to interpose a logistics service provider who coordinates the material flow streams of the individual sub-suppliers. (Hundertmark 2013, 18–19.)

All these disadvantages are accepted in return for order-specific and throughput-optimized production because the modules are not only delivered at the right time to the proper workstation for final assembly. Above all, they are delivered in the correct installation sequence, usually determined based on the customer orders. That makes in-house distribution even from intermediate buffers superfluous and improves the production flow to the highest degree.

Due to the almost instantaneous installation and the challenges of this delivery variant just described, it is common for the OEM to assume responsibility for process reliability from the supplier contractually. That applies at least for a few work cycles after the actual installation as well as in the event of a delivery failure. (Hummel 2019, 4–5.)

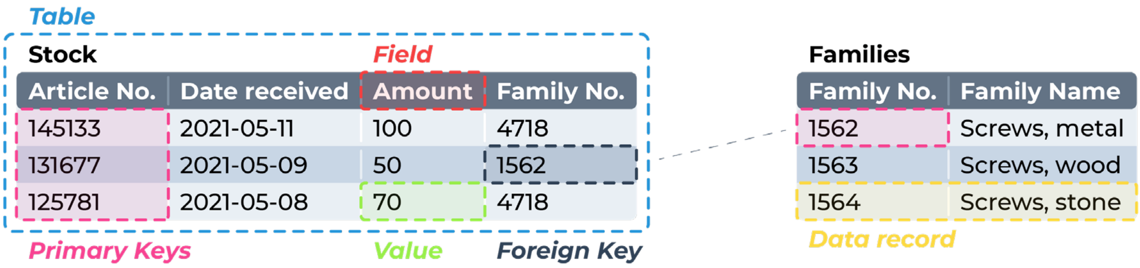
2.3 Data management and accessibility

2.3.1 Basics of databases and software development

When it comes to automation and digitalization, sooner or later, the use of one or more databases, i.e., organized collections of structured information (also called datasets), is almost inevitable. In contrast to conventional spreadsheet programs such as Microsoft Excel, they are particularly suitable for containing vast collections of organized information. Furthermore, they allow several users in parallel to call up data with complex relationships using queries easily. These queries are usually formulated using the structured query language (SQL), a programming language specially developed for databases. Databases have become indispensable today, especially for improving company performance and for decision-making. (Oracle n.d.a.)

A relational database refers to a collection of interconnected, two-dimensional tables containing and categorizing the corresponding data. As in any table, there are both rows and columns. The column names are called fields. Fields are always assigned a specific data type, which means a fixed default for the format of the contained values. The three most common data types are strings (a sequence of characters, whether numbers or letters), integers (whole numbers), and booleans (truth values, true or false). For example, if the data type integer is assigned to the table field 'article number', no value may be 'one hundred and forty-four' but only '144'. A complete row of a database table is called a data record or entry. While the number and types of table fields must be specified in advance, a table can contain any number of data records during its lifetime. (Steiner 2021, 5–6, 14.)

However, the individual values of the various tables of a database would be worthless if it were not possible to set relationships between them. For this purpose, each table must have a primary key by which a data record from this table can be unambiguously identified. Therefore, speech names such as 'screw', for example, would not be suitable as primary keys. There can be several types of screws in a database and several entities of the same type. Usually, defined number or ID fields are used as primary keys for this reason. In order to refer or connect to other tables, data records contain not only their primary key but also any number of foreign keys, which in turn are also primary keys of another table. In this way, information from different tables can ultimately be selectively linked and compiled via an SQL query to create new, individual datasets. (Steiner 2021, 15–16.) The terms just introduced are summarized in Figure 5 below.



Stock

Article No.	Date received	Amount	Family No.
145133	2021-05-11	100	4718
131677	2021-05-09	50	1562
125781	2021-05-08	70	4718

Families

Family No.	Family Name
1562	Screws, metal
1563	Screws, wood
1564	Screws, stone

FIGURE 5. Database terminology

The principle of least astonishment (PoLA) describes one of the primary areas of human-computer interaction (HCI). It refers to the development of software in general, mainly focusing on its interface. According to this principle and within the framework of the HCI, a perfect program is assumed that identifies people in their role as software users as the most critical point in the operating chain. Thus, the greatest risk of program malfunction is very likely the end user. According to PoLA, it must be ensured that the human user is neither surprised nor astonished by the program, as the resulting distraction can lead to incorrect operation. In other words, that means even if a more elegant or faster software solution was conceivable, a technically possibly more deficient solution should be preferred if it is less surprising for the human operator or more comprehensible to them. The PoLA is thus also part of the user experience (UX) and vice versa. (Yampolsky 2018.)

2.3.2 Lean user experience design

According to Lynch & Horton (2016), user experience (UX) describes a whole perspective from which the entire planning process and all tasks of web and app development must be considered. The questions 'How easy is the site to use?' or 'How efficient is the application?' and above all 'Did I find what I was looking for?' shall be asked. Usually, the quality of user experience is assessed according to how usable and enjoyable the application or page appears to people.

In order to ensure the best possible user experience, a distinction must be made between seven individual attributes, which in turn can be perfected individually. First of all, the learnability (1) contributes to an improved user experience. It assesses how easily and quickly a first-time user can learn to get the information or services they want from the application. The ease of orientation (2), on the other hand, describes how safely and precisely users can determine their location within the application's navigation system. The efficiency (3) is defined by the speed with which users can browse, search, or find the information they want. Memorability (4) is also essential, i.e., the ability to reactivate knowledge about operating the application after a long period of inactivity. Even users who have physical or sensory restrictions should be able to effectively use most of the application, which is expressed with the term accessibility (5). The error forgiveness (6) already touched on in the context of the HCI is once again of great importance. It is still important here to catch input and usage errors by human users and, of course, to avoid them in advance. Finally, as already mentioned, the application should also be enjoyable and not have to be seen as an annoying or even insurmountable challenge. That is described by the term delight (7). (Lynch & Horton 2016.)

Because the number of devices, users, possible applications, and thus also user wishes is growing steadily and at an ever-increasing rate, there is also a growing need to streamline not only the operation of an application but also its development process. Lean UX design is used in this context. (Hennecke 2015.)

As part of the lean UX design, the focus should be on developing only those functionalities that future users really need and appreciate. Therefore, this design concept largely overlaps with methods of agile project management such as Scrum management, which is explained in more detail in Chapter 2.4.2. Accordingly, the concept is primarily based on the development of lean prototypes created in collaboration with the users. Constant feedback loops from the beginning of the project are therefore essential for development. It is all about understanding potential users' problems and solving them instead of proactively developing features that may never be used. Measurable and comparable key performance indicators must also be introduced in order to be able to assess the development progress adequately. Tools and aids are used flexibly and only when they are actually needed instead of blindly following rulesets and procedures set in stone. (Hennecke 2015.)

2.4 Methodology

In addition to the theoretical background on logistics as well as data processing and management, this project also required practical information and methodologies independent of these. Crucial here is the project management, in the framework of which the tasks and goals were gradually worked through in a structured manner. Furthermore, the handling of large amount of data and the derivation of information from it is essential for the project's success. For these reasons, Scrum management as a variant of agile project management and structured data analysis will be examined in more detail below.

2.4.1 Structured data analysis

In order to analyze given data (sets), statistical methods are and must always be used. The desired result of such a (in the best case structured) data analysis is called information. This information obtained is intended to be used for decision-making and corporate planning, the actual goal of structured data analysis. A fundamental distinction is made between two variants of statistics, descriptive and inductive (or inferential) statistics. While descriptive statistics describe the data of an entire population to obtain information, inductive statistics take individual samples from which conclusions are then drawn about the population. (Kamps 2018.)

Nowadays, especially against the background of exponentially growing amounts of data, it is hardly possible to fall back on purely descriptive statistics despite the likewise rapidly increasing computing capacity. Rather, it requires an approach that makes use of the methods of both sides. In this so-called knowledge process, the source data is processed and evaluated in stages to form different bases of data. First, descriptive statistics are used as an at least in itself complete set of data is summarized and processed into information. Evaluation methods of inductive statistics are then used to gain generalizable knowledge from the newly acquired data basis, consequently influencing decisions and actions. (Cleff 2011, 3–6.) This step can also be found in the intelligence cycle described later and in Figure 6.

In particular, empirical research, i.e., planned observation and the contextualizing of data, is indispensable. This methodology is also carried out in several phases. First, there is the exploratory phase, during which, ideally, all possible cause-effect relationships are identified. Between the subsequent theory or model phase, feedback loops are quite important. Communication skills are required, and discussions with third parties and experts are conducted. That is the only way to find as many potentials or cause-effect relationships as possible. (Cleff 2011, 7–8.)

In the second, already mentioned theory or model phase, abstraction and simplification are required. The significant majority of all use cases are far too complex for them to be mapped in a completely identical manner compared to reality (i.e., isomorphic mapping). As a rule, homomorphic partial models are used to approximate reality. They are reduced in their complexity for better handling, only taking into account the most likely relevant factors. (Cleff 2011, 9–11.)

During this step, it is common to utilize the exploratory data analysis (EDA) for the abstracted data models. Primary targets are to isolate important variables, uncover underlying structures, and detect outliers as well as other anomalies.

Typical methods of EDA are plotting the raw data and simple statistics like box plots and residual plots. (Hinterberger 2009.)

In this phase of the structured data analysis, the homomorphic partial model, consisting of methods of descriptive and inductive statistics, can be transferred to the intelligence cycle (Figure 6). However, the target group of the information to be determined must always be taken into account in this step. The original goals of statistics, decision-making and corporate planning, cannot be achieved if the relevant decision-maker does not understand the generalizable knowledge created. Therefore, analyses and results must always be designed so that the information needs of management are met. The intelligence cycle closes when the decisions made based on the generalizable knowledge generate new results, which in turn form the data basis for the renewed generation of information. (Cleff 2011, 13–14.)

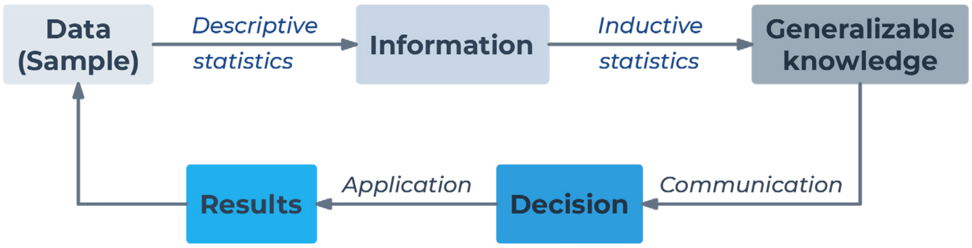


FIGURE 6. The intelligence cycle (Cleff 2011, 14, modified)

In the following phases, the investigation and the evaluation phase, it is a matter of checking the correctness of the cause-and-effect principles brought into a model using a wide variety of methods. Which method should be used depends mainly on the type of data as well as the desired level of contextualization. (Cleff 2011, 14.)

The IT-based intelligence cycle and information derivation process, especially for business use cases, has its own name, business intelligence (BI). Therefore, BI tools and their aid for decision making grow in popularity. Of course, without connection to digital databases these tools are not able to function. (Siepermann & Lackes n.d.)

2.4.2 Project planning and management

Especially in software and application development, not all project contents are necessarily known at the beginning of the project. Only the requirements or desired functions can be defined, but usually, the customers or the users cannot formulate these goals themselves clearly. That leads to a dynamic project environment, non-binding planning, and makes changes more likely. Accordingly, a simple transfer from the product requirements document to the duties record book and its subsequent linear processing is hardly conceivable. Implementing changes in requirements in classic project management becomes exponentially more expensive as the project time passes. (Fleig 2019.)

Against this problem's background, it has proven useful to stay in permanent contact with the customer and give them perpetual feedback while requesting changes and calculating constant change costs. These are the basics of agile project management.

One of the best-known variants of agile project management is referred to by the term 'Scrum'. Essential for this methodology is an iterative, i.e., step-by-step development of functional sub-products (the so-called product increments) presented to the customer in reviews that are to be held regularly. The project team should organize itself in cooperation with experts and developers, assume a high level of responsibility for the product, view changes as standard, and use them as opportunities. In addition, only the work is carried out that is considered to be absolutely necessary for reaching the next product increment. (Fleig 2019.)

A simplified form of Scrum was mainly used to work on this project, which is why the most important terms and procedures are briefly discussed below. The three main roles must be determined in advance:

- The development team that is responsible for the implementation of the defined tasks
- The Scrum master, also known as the coach who exclusively monitors the methodology, its problems, and correct application

- The product owner, the interface to the customer who defines, prioritizes, and documents the tasks in the so-called product backlog

As soon as the product backlog is filled, an intermediate goal prioritized by the product owner, a sprint goal, is selected from this, which is then usually pursued by the development team within two to four weeks. At the same time, the basis and the goal of a sprint are always the product increments that are presented to the customer and serve to adjust the product backlog if necessary.

A sprint planning meeting precedes such a sprint, in which the product owner presents the sprint goal, and the development team decides how and at what time this goal can be achieved. The meeting's output is a list of tasks containing the functions necessary to achieve the goal (sprint backlog). (Jungwirth 2016.)

During the daily Scrum meetings, the current processing status of the sprint backlog is constantly audited, e.g., using a burndown chart or Gantt chart, and the tasks for the next 24 hours are decided and distributed. Once a sprint has been completed, two 'inspect-and-adapt' activities follow. These are the sprint review, which is used to assess the actual product, and the sprint retrospective, which is used to evaluate the development process. (Jungwirth 2016.)

In addition to the Gantt chart, another method for task visualization was used during this project. The Scrum board offers a clear option for monitoring the current project progress and the product backlog points (PBPs) still to be completed. The individual items of the product or sprint backlog are placed on a board, which is divided into three sections. On the far left, all items still to be processed ('To Do') are located. In the middle, all items currently being worked on are pinned ('In Progress'). Finally, on the board's right side, the completed tasks are displayed ('Done'). The goal is, of course, that at the end of a project phase, all tasks can be moved to the 'Done' column. (van der Wardt n.d.)

3 PRACTICAL IMPLEMENTATION

3.1 Volkswagen AG and digitalization strategies

Volkswagen Aktiengesellschaft (Volkswagen AG) is based in Wolfsburg, Lower Saxony, Germany, is one of the world's leading automobile manufacturers and the largest car manufacturer in Europe. The group operates 118 production facilities in 30 countries, employs over 660,000 people, and sells vehicles in 153 countries (Volkswagen AG n.d.d). The most significant competitors on the world market are presently Toyota and General Motors (Statista 2021a).

Through far-reaching partnerships as well as start-up cooperation and foundation, Volkswagen AG constantly expands its competencies and areas of activity to achieve its self-declared business goal. This is, delivering sustainable, individual, and affordable mobility through cleaner, quieter, more intelligent, and safer automobiles and thus with electric drive, digital networking, and autonomous driving. (Volkswagen AG n.d.d.)

Of great relevance to this project are the recently concluded cloud collaborations with Microsoft (since 2018) and Amazon (since 2019). On the one hand, the Volkswagen Automotive Cloud is being developed based on the Microsoft Azure cloud. The focus of this cloud is autonomous driving and thus the networking of vehicles with each other and with the customer. The aim is to develop a digital ecosystem that can be integrated into the customer's everyday life, which is ultimately intended to transport customers fully autonomously from a specified starting point to the desired destination. (Volkswagen AG 2019.)

An expansion of the core of the Automotive Cloud partnership, the Automated Driving Platform, was only announced in February 2021 (Hüttel 2019). With the establishment of the CARIAD organization by Volkswagen AG, development using Microsoft Azure is to be simplified, and Volkswagen's own share in the development of automated driving functions is to be increased (Microsoft Corporation 2021).

On the other hand, the Volkswagen Industrial Cloud focuses on the production side of the group (Haak 2019). It is intended to represent the essential technological prerequisite for achieving the productivity goals in production by bringing together the data of all machines, plants, and systems from all of the group's factories. In the long term, data from the global supply chain from more than 1,500 suppliers and partner companies should also be made available and evaluable in the Industrial Cloud. (Volkswagen AG 2020b.)

3.1.1 Logistics planning of Volkswagen AG Commercial Vehicles

Volkswagen Commercial Vehicles is one of the twelve group brands of Volkswagen AG, currently employs around 24,000 people worldwide, and focuses on the production of light commercial vehicles. At the main location in Hanover, currently, the Transporter series (also known as the "Bulli") is manufactured, and the electrically powered large transporter e-Crafter is processed. Furthermore, the regular Crafter and the light van Caddy are manufactured in Poland. (Volkswagen AG 2021c.) An overview of the VW CV Europe model range is shown in Picture 1. From left to right the Caddy, Transporter 6.1 (as a California model), and the Crafter (as a Grand California model) can be seen.

PICTURE 1. The VW CV model range in Europe (Volkswagen AG 2021b)

As an automobile manufacturer and OEM, VW CV has also achieved supply chain management. That means that a large part of the warehouse space has been gradually converted into production space, while just-in-time and just-in-sequence delivery concepts have been established. In addition, material and flow of goods as well as development-related services were transferred to these very same suppliers against the background of an intensified relationship with first-tier suppliers – with all the consequences described in Chapter 2.2.1. Nonetheless, all four delivery variants are still used. Appropriate procedures have proven their worth depending on the module, component, or individual part. Here, too, the trend is clearly towards increasing modularization and integration of the suppliers. [redacted]. (Volkswagen AG 2020a.)

In contrast to the classic disposition, which is responsible for the direct supply of production with the necessary resources, i.e., the individual parts, components, and modules, logistics planning does not take on executive but mainly conceptual tasks. For VW CV, this means that the material flow planning for the producing part of the plant, i.e., the press shop, body shop, paint shop, and assembly, takes place here. In close cooperation with procurement, specifications for supplier contracts are drawn up, and delivery options for individual parts are determined. In addition, the planning for logistics and distribution centers, material flow collection points, and other storage areas, both inside and outside the plant, is done here.

3.1.2 Amazon Web Services and the Logistics Data Lake

As part of the Volkswagen Industrial Cloud group strategy (Chapter 3.1), all logistics data for the entire Volkswagen Group are to be stored and made accessible centrally. As already mentioned, the Amazon Web Services architecture is used for this. Based on this platform, which is tailored to the group's needs, all further development steps are taken to achieve the overarching goal of increasing productivity in all production areas. (Volkswagen AG 2019.)

Various products from Amazon Web Services (AWS) cover the numerous requirements for the software architecture of such a central system. Amazon Web Services, a subsidiary of the online mail-order company Amazon, is one of the leading cloud computing providers on the world market. Furthermore, the pricing of the services offered is based on their actual use. Volkswagen mainly uses the PaaS variant of cloud computing. Consequently, AWS provides the platforms and their maintenance, but Volkswagen can still make programmatic changes to adjust their functionalities. That means, for example, that queries to databases are not necessarily unchangeable or can only be put together via the user interface of the corresponding query interface (Amazon Web Services n.d.e). A complete depiction of the AWS technologies used in this project can be found in Figure 7.

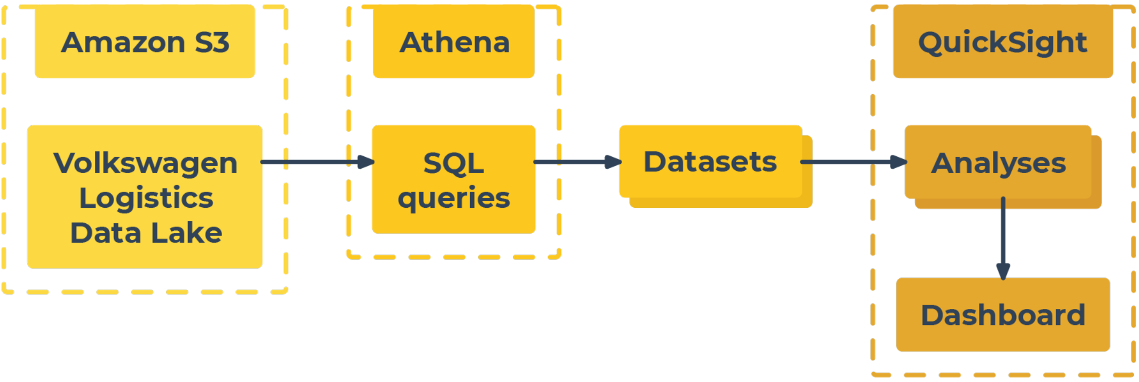


FIGURE 7. Dataflow and applied AWS technologies.

The migration of own computer centers and servers to the architecture of AWS enables completely new application variants for the storage and analysis of the logistics data of the group. Large amounts of data can be managed, grouped, and evaluated significantly faster using the big data analysis tools provided. When analyzing the data, it must be considered that the data structure and histories have also changed with the migration. AWS also takes care of the essential areas of cyber security and compliance requirements, all of these in combination with Volkswagen's internal Public Key Infrastructure (PKI) identification system.

The PKI system is based on two types of keys. First, the private (secret) keys, and second, the public keys. The possession of a private key alone enables the decryption of information that has been encrypted with the corresponding public key. The so-called registration authorities (RAs) and certification authorities (CAs) are required to issue and manage private and public keys. The RAs are physical contact points that check the applicant's identity and, after confirmation, can make requests to the CAs for specific certificates. A certificate is a digital attestation for the applicant and his private key about the possession of a specific public key. This certificate is issued by the electronic signature of a CA. The CAs themselves are represented by certificates, which in turn are signed by a hierarchically higher CA. That results in a CA hierarchy, at the end of which is the root CA, whose certificate is signed by itself. After successfully generating a certificate for a user and a specific public key, this is registered on the personalized Volkswagen PKI card. Therefore, in this use case, the data and tools of Volkswagen Amazon Web Services only can be accessed with the corresponding PKI card (which is read via a card reader in the operating device, usually in the corporate laptop). Also, a successfully registered matching certificate in conjunction with the user's private key is needed. Moreover, the certificate, together with the private key, has to be sent to the access point in refresh at least every 60 minutes in order to prevent abuse. In this way, it is even possible to directly trace which user made changes where and when in case of doubt. (Volkswagen AG n.d.a.)

A data lake with Amazon S3, the Amazon Simple Storage Service, is currently being set up to store all logistics data. According to AWS, S3 is designed to provide data access in 99.999999999 percent of the cases (Amazon Web Services n.d.d). In addition, precisely coordinated access controls can be configured via S3, preventing unwanted access from external parties. Accordingly, there are currently three different roles for the Volkswagen Logistics Data Lake. The 'business analyst' who can use basic tools for analyzing existing datasets out of security level 'internal information', the 'business analyst confidential' who has the same tools available like the regular business analyst but can access datasets out of security level 'confidential information' in addition, and the 'data scientist' who is able to access all data and create new datasets as described below. (Volkswagen AG 2021a.)

Data scientists can use Amazon Athena, the interactive AWS query service, to query data directly from S3. Athena is a serverless system, which means that there is no need to extract, transform and load the data. All that has to be done is setting up a reference to the desired data from S3 and defining a schema so queries can be sent directly afterward using SQL. In Athena, too, payment is only made for a query that is actually carried out (5 USD per analyzed terabyte), which usually saves 30% to 90% of the query costs compared to the internal, company-owned server center (Amazon Web Services n.d.a). Athena is based on the SQL dialect Presto, a high-performance SQL specially developed for big data analyses. The advantage here is that several data sources can be accessed simultaneously in a single query, which significantly streamlines query programming. The queries themselves are further accelerated because Athena can execute the various queries in several locations and on several devices per location concurrently. (Amazon Web Services n.d.a.)

As soon as the queries that have been designed with Athena behave in a stable manner and deliver the desired datasets, the queries and datasets can be transferred to Amazon QuickSight. QuickSight is a scalable, serverless, and machine learning-based business intelligence (BI) service developed explicitly for the cloud. With QuickSight, interactive BI dashboards can be created to create abstracted and quickly comprehensible data representations and analyses according to the intelligence cycle. The main advantage of QuickSight is that once a dashboard has been created, it can be released immediately for an almost unlimited number of other users. In addition, the queries generating datasets to be analyzed in the dashboard can be automated so that the data basis is always up to date. (Amazon Web Services n.d.c.) QuickSight also offers machine learning elements so that after some time, forecasts can also be created based on data records queried in the past (Amazon Web Services n.d.b). Query integration can also be done with Amazon SPICE, an in-memory computing engine, which accelerates the updating and compiling of the datasets. (Amazon Web Services n.d.c.)

3.1.3 Purchasing and distribution centers as data sources

In this chapter, the data generation process, and data specifications of the Volkswagen Logistics Data Lake (LDL) shall be explained. Since VW CV already produces primarily according to the just-in-sequence concept, hardly any warehouses in the conventional sense are required anymore. However, there is a need for space for sequencing, grouping, and checking the modules to be delivered for production. Despite production planning that is accurate to the minute, deliveries for one of the three shifts of a weekday can arrive delayed due to unfavorable environmental conditions such as traffic jams or bad weather. In addition, various types of modules such as cockpits, windowpanes, or side mirrors are delivered by numerous suppliers from different parts of the world.

For this purpose, the purchasing and distribution centers (PDCs) exist outside the actual manufacturing plant as an interim storage, where the consignments for the next few hours of production are delivered. The parts get conveyed by driverless transport vehicles over a bridge transport route to the requesting points in production when needed. The two PDCs from VW CV are currently not operated by the company itself but by external logistics service providers. In this case by the companies [redacted] and [redacted] In turn, these service providers are in contact with the first-tier suppliers regarding the time of goods receipt on behalf of VW CV, coordinate the storage areas and employ the staff in the PDCs. Only after crossing the bridge transport route the responsibility and liability for the delivered parts is transferred back to VW CV again.

The PDCs and their content can be understood as a cyber-physical system. Several objects contribute to the creation of a data record. The exact structure of such a data record is discussed in more detail in Chapter 3.2.3. However, the containers in which the required parts are delivered are standardized and integrated into a return cycle. Each container is provided with a barcode that mobile reading devices can scan. Different data is generated depending on where this reader is located. The so-called material flow status is then derived from this. A distinction is made between six different statuses. Material on transporter ('1'), on plant premises ('2'), at the goods receipt area ('3'), successfully stored ('4'), conveyed to consumption in production ('5') and conveyed to shipping ('6'). Only those parts with a material flow status of '4', stored, are relevant for evaluating the warehouse utilization.

For the following calculation methods and the analysis of conceivable logical cases, it is necessary to differentiate the term container further. In the PDC, modules required for production are delivered on various load carriers. The smallest possible unit is the package. A package is always a single load carrier with defined dimensions. A package ultimately contains the individual modules directly, such as X rear bumpers. A package is always assigned to a container number.

In turn, a container can contain one or more packages. Frequently, several packages with the same contents are delivered on a pallet. This bundle of pallet and packages is therefore also called a container. Accordingly, only the dimensions of a container are relevant for the calculation of the warehouse utilization. The different variants of a container and container types are shown in Figure 8.

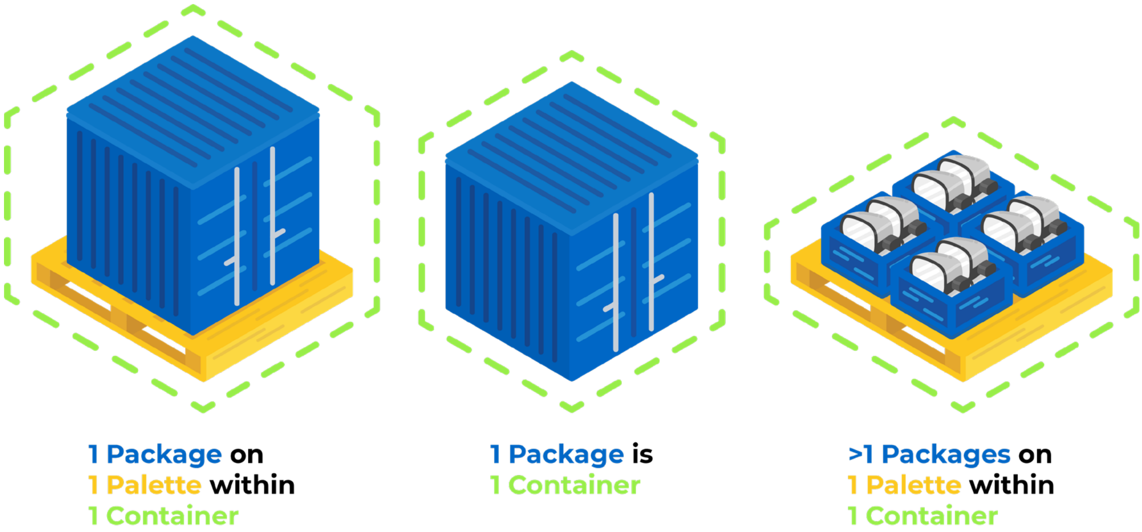


FIGURE 8. Occurrence possibilities of a container

The other part of the communication between components takes place through the storage location itself. The PDCs contain various storage location types, including block storage (reserved, rectangular areas for stacking containers), aisles (reserved rows for stacking containers), racks, and an automatic miniload system (AMS), which can automatically store and retrieve smaller containers in inner rack aisles. When a part is stored, e.g., via a forklift truck, a data record is generated again, containing information about the container's storage. Therefore, a finished data record contains the information about the container and what it contains (data origin A, from the container itself), the material flow status of the container (data origin B, from the transport vehicle/reading device), and where it is stored (data origin C, from the storage location). This communicative context can also be found in Figure 9. This data is stored in different systems or, rather, different databases. One of the hurdles during the project implementation was to find out where, when, to what extent, and with which key this data was stored.

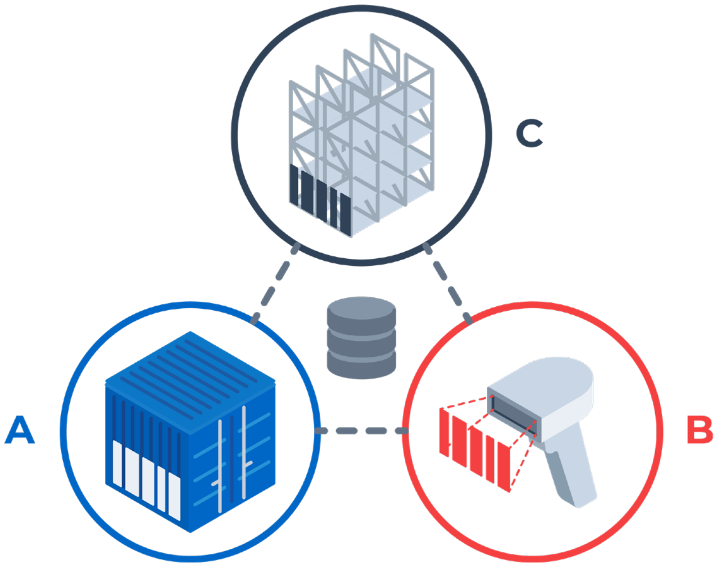


FIGURE 9. PDC as CPS, communication instances

3.1.4 Problem description and use case

In the past, the question has arisen why VW CV needs the utilization figures of their PDCs when these are managed by external service providers such as the [redacted] and [redacted] The objection is justified since the mentioned companies carry out their operative business largely independently. However, the core of the necessity of a utilization analysis can also be found here. The service providers only carry out operational steps. The planning, optimization, and adaptation of the just-in-sequence system are still in the hands of VW CV. Area extensions, rack planning, and change management are always designed in the process management department of the logistics strategics and then get implemented in coordination with the service providers.

In addition, the daily guarantee of production security and stability for VW CV in their main plant adjacent to the PDCs naturally depends on the availability and predictability of the PDCs and their content. In addition, the production program planning must be shared with the service providers in order to then jointly determine needs for the coming production shifts, taking into account the remaining stocks. That is why the service providers and VW CV also share common databases in which this information can be exchanged.

A decisive key performance indicator (KPI) for the security of supply of the PDCs is, of course, their utilization in percent. This KPI is even more critical for calculating the average capital tie-up costs and the capacity of agreed deliveries planned in the future. Roughly speaking, the busier the PDC is (usually from >80%), the more difficult it is to provide the correct parts needed in production. The warehouse is referred to as "rigid" when the utilization is increased. That means parts that may be needed for production are stored further back in aisles, or more dramatically, further down in stacks in the block storage areas. Consequently, far-reaching relocation processes are necessary. In addition, there is less space for those relocation processes due to excessive utilization, which also leads to delays or overtime.

On the other hand, and from VW CV's point of view, the capacity utilization of a PDC should not be too low in the long run since otherwise too much rent for unused space or salary (through the service provider) for staff that is not needed will be paid. Moreover, a too low utilization can also indicate that the required parts could not be available in the prescribed flexibility and amount. For these reasons, there is an optimal utilization for PDCs, with which time sufficient security of supply is guaranteed, but also adequate delivery flexibility is available at the same time. It is essential to keep to this utilization as consistently as possible. That is exactly why it is so crucial to monitor this KPI carefully.

The utilization in percent is ultimately only a construct to approximate the properties mentioned above, such as security of supply. Therefore, it is a very highly abstracted homomorphic partial model. Of course, more detailed insights are necessary for the exact determination of needs. Nevertheless, no human can process and evaluate the raw data that is available in a meaningful way. Therefore, further partial models have to be developed, which in turn concentrate on specific areas of interest. Questions that are asked against this background are: How many type X containers are stored at the moment? Where exactly are these containers located? To what percentage is rack aisle X being used? And most importantly, could certain relocation processes reduce the utilization of the warehouse?

3.2 Development of an optimized utilization analysis

A solution now had to be developed to answer the questions posed above. Therefore, this chapter deals with the structured development of a data analysis tool designed specifically for this task. In iterations within the Scrum management framework, an evaluable, suitable dataset was created from the entire LDL dataset via Amazon Athena (step 1 of the intelligence cycle, descriptive statistics). This dataset is then transferred to Amazon QuickSight to derive generalizable information based on this homomorphic partial model (step 2 of the intelligence cycle, inductive statistics) and to present this information appropriately for decision-makers in accordance with the principles of lean UX design.

3.2.1 Project initialization

Because this is an IT-oriented project aiming to merge and evaluate data, it was decided in advance to carry out a variant of project management based on Scrum. An initial project plan as a Gantt chart was created in the very beginning. It can be found in Appendix 2. However, it has been changed and separated into four three-week Scrum sprints later during the project, which are described from Chapter 3.2.3 on. After the rough chart was created, the functions of the three main Scrum roles have been assigned as follows.

TABLE 1. Person-function assignment of the Scrum roles

Person	Functions taken over of the Scrum roles
Subdivision manager	Product owner | Prioritizing sprint goals
Supervisor in subdivision	Product owner + Scrum master | Creating the product backlog, daily Scrum, sprint reviews
Student	Scrum master + development | Scrum initialization, sprint execution, daily Scrum, sprint retrospective

After the target questions shown in the previous chapter have been formulated, the already existing solution approaches must be considered first in order to be then able to put together a product backlog. At the start of the project, two methods that build on each other already existed for the PDC utilization analysis, but both had significant disadvantages. The product backlog was created based on the functionalities that already meet the requirements as well as the missing features for the correct utilization analysis as described below.

Since the individual items of the created product backlog were not sufficiently separatable into sub-tasks, it was decided to design the Scrum board based on the individual PBPs. A sprint backlog is therefore always visible in the 'In Progress' column. Overall, the PBP processing was divided into four consecutive three-week sprints, as shown in the following chapters.

3.2.2 Requirement analysis and product backlog

SAP Business Objects evaluation

The original and therefore older evaluation method for the PDCs is carried out as a database query using SAP Business Objects (BO). As already announced, however, this BO evaluation has weaknesses. Above all, the KPI itself is the problem. The utilization (in the following referred to as κ) is not calculated based on the space available in the PDC but based on an average number of containers that could possibly find space in the warehouse. This calculation method assumes that all parts – be it door locks or windowpanes – are delivered in exactly the same container with the same dimensions (height h_{hyp}, width w_{hyp}, depth d_{hyp}). Then, with the help of these dimensions, it is calculated how many of the hypothetical standard containers would fit into the warehouse ($n_{max,hyp}$). This number then gets compared with the number of containers actually booked in the warehouse (n_{stored}) – regardless of their type (see Formulas 1 and 2).

$$n_{max,hyp} = \frac{total\ storage\ volume\ available}{h_{hyp} \cdot w_{hyp} \cdot d_{hyp}} \tag{1}$$

$$\kappa = \frac{n_{stored}}{n_{max,hyp}} \tag{2}$$

The problem is that, of course, not all parts are delivered in the same container. Despite standardization, containers and pallets have very different dimensions and cannot be offset against each other in this way. The utilization indicator, which results from the BO evaluation, therefore is only used as a rough guide but, unfortunately, never reflects the exact utilization.

A significant advantage of this BO analysis for the subsequent project implementation is that at least in the old databases that have not yet been migrated to S3, the currently stored packages are correctly identified via the so-called package reference number. However, there are deficits here, too. The packages are correctly recorded based on their storage location (data origin C), their material flow status (data origin B), and their number (data origin A, partially). However, the individual data record is not always correct since in the case of one or more containers stored on a pallet, the wrong dimensions, namely those of the pallet, are pulled from the database.

Advanced spreadsheet calculation

Since κ from the BO analysis is not reliable enough for the reasons mentioned, another method was developed to get as close as possible to the actual utilization of the PDC. It is based on the dataset obtained from the various databases as part of the BO evaluation. However, the calculation of κ is much more detailed. A large part of the first phase of project implementation deals with analyzing this calculation algorithm using the process available in the spreadsheet program Microsoft Excel.

First, the queried data is re-segmented. A distinction is made between the individual racks, aisles, and blocks. Each of these elements is assigned a maximum capacity for each possible container type ($n_{max,e,t}$). Thus, it is assumed that only a single container type t is stored in the corresponding storage element e (rack, aisle, or block). The actual dimensions of the package (height h_t, width w_t, depth d_t) that lies in a specific type t container and the stacking factor s_t of a container are now taken into account. The stacking factor indicates for each container type how often it can be stacked on entities of its own type. For the calculation of $n_{max,e,t}$, however, a further distinction is made between aisle and block ($n_{max,ab,t}$, Formula 3) and rack ($n_{max,r,t}$), where of course ab and r are both e elements. The calculations are as follows with l_{ab} as the length of an aisle or block and x_t as the container dimension in line with the storage element length (could be the width

w_t or *d_t*). For *$n_{max,r,t}$*, the calculation (Formula 4) is a little more complex as in addition to the width of a rack compartment *w_{rc}*, the number of rack compartments next to each other in one rack *$n_{rc,rl}$*, and the heights of each of the four rack levels *h_{rl}* have to be considered.

$$n_{max,ab,t} = \frac{l_{ab}}{x_t} \cdot s_t \tag{3}$$

$$n_{max,r,t} = \left\lfloor \frac{w_{rc}}{x_t} \right\rfloor \cdot n_{rc,rl} \cdot \sum\nolimits_{rl=1}^{4} \left\lfloor \frac{h_{rl}}{h_t} \right\rfloor \tag{4}$$

with

$$n_{max,ab,t}, n_{max,r,t} \subseteq n_{max,e,t} \tag{5}$$

The current utilization per storage element and container type *$n_{e,t}$* is then compared with the individual maximum capacities. That results in partial utilizations (*$\kappa_{e,t}$*) which are added up for each storage element and existing container type (set *T*). That leads to the utilization per rack, aisle, or block (*κ_e*), shown in Formulas 6 and 7.

$$\kappa_{e,t} = \frac{n_{e,t}}{n_{max,e,t}} \tag{6}$$

$$\kappa_e = \sum_{t \in T} \kappa_{e,t} \tag{7}$$

However, the last step in calculating *κ* is again based on the number of pieces. The utilization of the individual storage elements *κ_e* is averaged based on their occupancy in pieces (*n_e*) in relation to the total amount of packages booked in the warehouse (*n_{total}*) to obtain the total utilization *κ*. Thereby, *E* is the set of all storage elements existing:

$$\kappa = \sum\nolimits_{e \in E} (\kappa_e \cdot \frac{n_e}{n_{total}}) = \sum\nolimits_{e \in E} (\frac{n_e}{n_{total}} \cdot \sum\nolimits_{t \in T} (\frac{n_{e,t}}{n_{max,e,t}})) \tag{8}$$

Although this method comes closer to the reality of actual warehouse utilization, it also brings computational as well as practical problems with it. At first, it is not automated. The dataset from BO must always be loaded manually into Excel. That means the evaluation is not available immediately. Two other factors also influence the quality of the evaluation. On the one hand, the final calculation of utilization κ (Formula 8) again is based on the number of items and not on the storage volume. On the other hand (and having a much more severe effect), during the calculation of the maximum capacity per storage element and container ($n_{max,e,t}$), it is assumed only containers of one type are stored in a storage element (Formulas 3–5). As a result of this simplification, when adding up the individual utilizations, utilizations over 100% are sometimes created. For example, although only 3 large containers of type t_1 or 6 small type t_2 containers would fit, with a clever arrangement shown in Figure 10, 2 large t_1 containers and 3 small t_2 containers can possibly be stored as well in one storage element. According to Formula 8, this would then lead to a utilization of 116.6%.

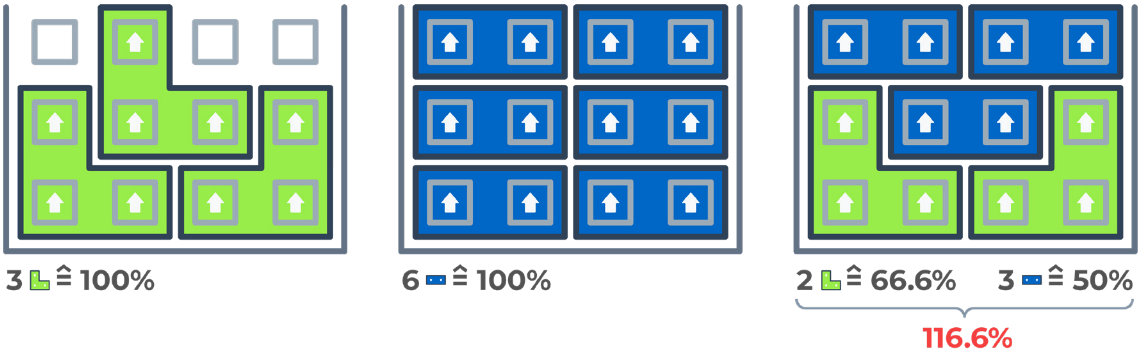


FIGURE 10. Compression effect in an aisle storage

The last problem with this calculation lies in the quality of the data basis and is related to the definition of a container. If it is a container in which a pallet serves as a load carrier, mistakenly not the dimensions of the entire container (pallet plus packages) are used for the calculation but only those of the pallet. That leads to completely excessive individual maximum capacities because pallets can naturally be stacked on top of each other many times more than other load carriers that are significantly higher in comparison.

In accordance with the previous calculation methods' advantages and disadvantages described, the product backlog was then created. A tool must be developed for Volkswagen AG that has the following functionalities:

1. Utilization analysis of any Volkswagen PDC, including:
    a. Total utilization of a PDC
    b. Utilization of individual storage areas and storage elements
    c. Distribution of the different container types
2. Therefore, identification of required logistics data within the Volkswagen Logistic Data Lake, connection establishment, and dataset creation
3. Regular, automatic updates of the datasets to be evaluated
4. Immediate availability of the evaluation analysis
5. No calculation based on the assumption of average dimensions or single-type storage, but based on the actual dimensions of containers and the available storage volume (parametrization of calculation)
6. No number-based but exact volume-based calculation of utilization
7. Straightforward graphical traceability of the processed evaluation with the possibility of more detailed insights
8. Forecast of future utilizations based on past utilizations and container type distribution

3.2.3 Data identification

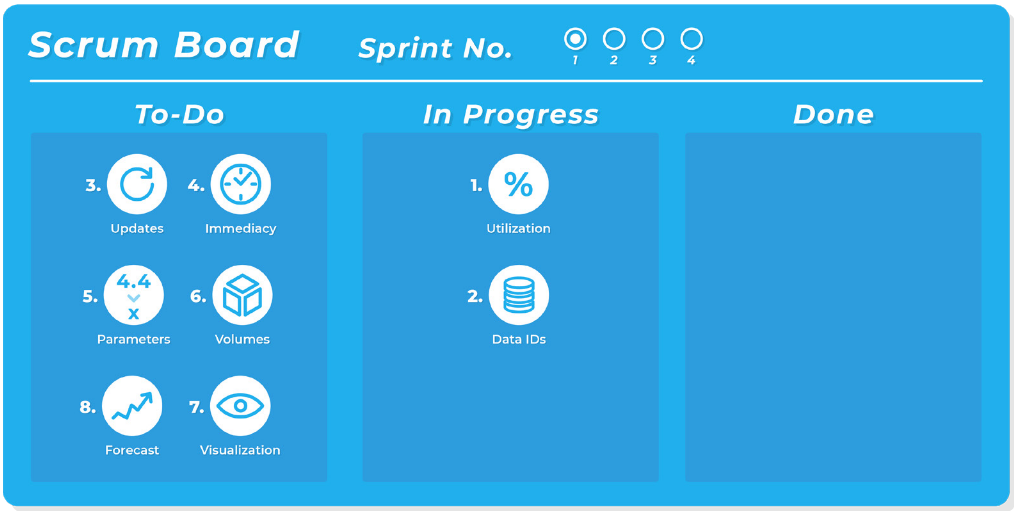


FIGURE 11. Scrum board at sprint 1

Figure 11 shows the developed Scrum board according to the now determined PBPs at the beginning of sprint 1. For the reasons listed in Chapter 3.2.2, it was decided according to the corporate strategy not to develop the now concretized tool using SAP Business Objects but rather based on the Amazon Web Services architecture. After the list of functionalities had been created, the first sprint could be started forthwith. This sprint deals purely with the processing of PBP 2, i.e., with the identification of the data needed for the information required for PBP 1 in the Volkswagen Logistics Data Lake. First, a data record scheme was created in which all the core values needed to achieve the goal are listed. These can be seen in Table 2.

TABLE 2. Core data record values

Information	Use / determination	Abbreviation
Package reference number	Unambiguous identification, counting	*PRN*
Material flow status	Currently stored package?	*mflstat*
Container reference number	Unambiguous identification, counting	*CRN*
Container type	Type distribution	*t*
Container dimensions (height, width, depth)	Single storage element utilization	h_t, w_t, d_t
Storage area	Location	*SA*
Storage place/element	Location, single storage element utilization	*e*
e dimensions (height, width, depth)	Single storage element utilization	h_e, w_e, d_e
Update date	How new is the data record?	*update*

The above-shown data would be sufficient for the pure calculation of the PDC utilization. However, in the LDL, there is no database in which this data is simply available for every PRN present. This information must be accessed through different access points. In the past of Volkswagen AG, numerous databases were set up that were intended to be used for the administration and evaluation of logistically relevant information. However, since similar but not exactly the same information was often required at different points within the company, the databases sometimes overlapped in their application areas. With the migration of the databases to S3 in the LDL, these ambiguities could be prevented from the outset. The database names were retained, but the table structure within them has fundamentally changed. That made a simple transfer of the previous SQL queries from SAP BO into Amazon Athena impossible.

Two target systems were identified in which the data required in Table 2 are available. The most important component here is [redacted] ([redacted]). The system is used to record material flow movements and control the material flow, bundling all information on incoming purchased parts as well as on in-house production materials for the areas of material receipt and inventory management. (Volkswagen AG n.d.c.) The associated [redacted] S3 database currently contains [redacted] tables, which needed to be checked for their relevance to the PDC evaluation. The second target system for tool development is [redacted] (Ger.: [redacted], Eng.: [redacted]). In turn, this system contains information about container master data, packaging planning, and empty container order processing. (Volkswagen AG n.d.b.) In concrete terms, this is where the container dimensions will be found. The associated S3 database currently contains [redacted] tables, of which only two will be used in the further course of the project.

Operational data store vs. storage layers

Due to the already mentioned restructuring of the tables, new links between individual tables have become necessary. An essential difference, which is also responsible for new identification criteria, is the data freshness, i.e., topicality of the individual data records. In the SAP BO evaluation used so far, the primary data source is an ODS table from [redacted] PACKAGE_ODS. ODS stands for operational data store. As can already be deduced from the term, this is short-term storage for the package data. The advantage of an ODS is that only the actively stored packages are available in the table. The material flow status is therefore irrelevant against this background. All package reference numbers present in PACKAGE_ODS are also currently stored in the warehouse. However, logically, the problem with this evaluation is that it represents only a snapshot at the exact time of the query. A history or prediction for the warehouse utilization cannot be realized utilizing ODS. For this reason, data from the storage layers (SL) are accessed for the development of the new tool concerning PBP 8. The table PACKAGE_SL contains all package reference numbers that have ever been registered.

The change from ODS to SL has the following consequences. First, the history of a package is completely traceable. At each point in the material flow of a package ('1'–'5', see Chapter 3.1.3), a new data record is generated for a PRN. This process causes, that a package in conjunction with its location can no longer be uniquely identified by a PRN. Since all and not only the current PRN are contained in the SL, it does not make sense from a storage capacity point of view to store the container dimensions and container assignment or container reference number (CRN) for each PRN here as well. A second table in [redacted] SL, `CONTAINER_SL`, stores all container reference numbers that have ever existed and at the same time provides information about the container type *t* of a CRN. Thus, it also determines whether the container is a pallet with (possibly several) packages on it or a single package. A third table, `LINK_SL`, is used to assign a CRN to each PRN. A connection in `LINK_SL` is in turn assigned a fixed start and expiration date. That makes the PRN reusable. For this reason, a new key is assigned to the combination of PRN, CRN, and the validation dates. This is the package data warehouse ID (PDWHID).

Three additional tables are still needed for the completion of an ideal data record from Table 2. A simple connection could be made to `STORAGEPLACE_SL`, a [redacted] table containing the dimensions (h_{SP}, w_{SP}, d_{SP}) of a storage element *e*. The correct container dimensions h_t, w_t, and d_t have to be taken from the so-called packaging datasheets. This, in turn, is [redacted] information, which is why a more complex identification requires four different keys. These are the part number (PrtN), the plant number (PlntN), the supplier number (SpplN), and the usage indicator (UID). The part number refers to the parts actually contained in the package, such as exterior mirrors. The plant number indicates which plant requests the package, in this case, always Hanover. The supplier number identifies the corresponding supplier in the LDL. The usage indicator is for differentiation between purchased and self-produced parts. An overview of all numbers and IDs can be found in Figure 12. Via the shown keys, an unambiguous connection to the tables `PACKAGING_SHEET_HEAD` and `PACKAGING_SHEET_BODY` is established, which have stored the container dimensions. A representation of all tables and data record origins is Figure 13.

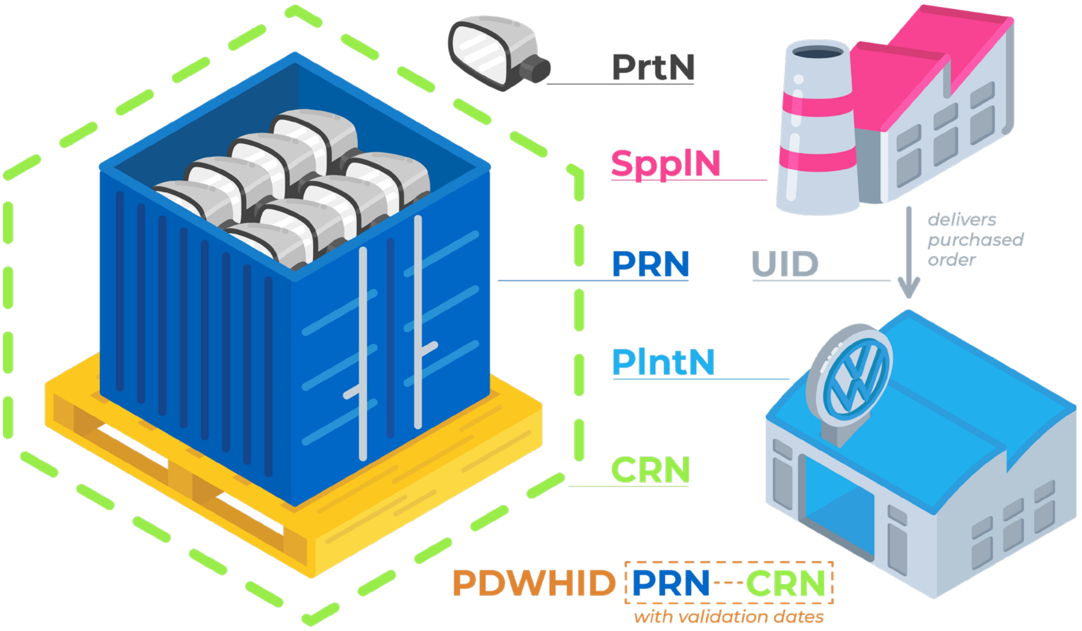


FIGURE 12. Identification numbers overview

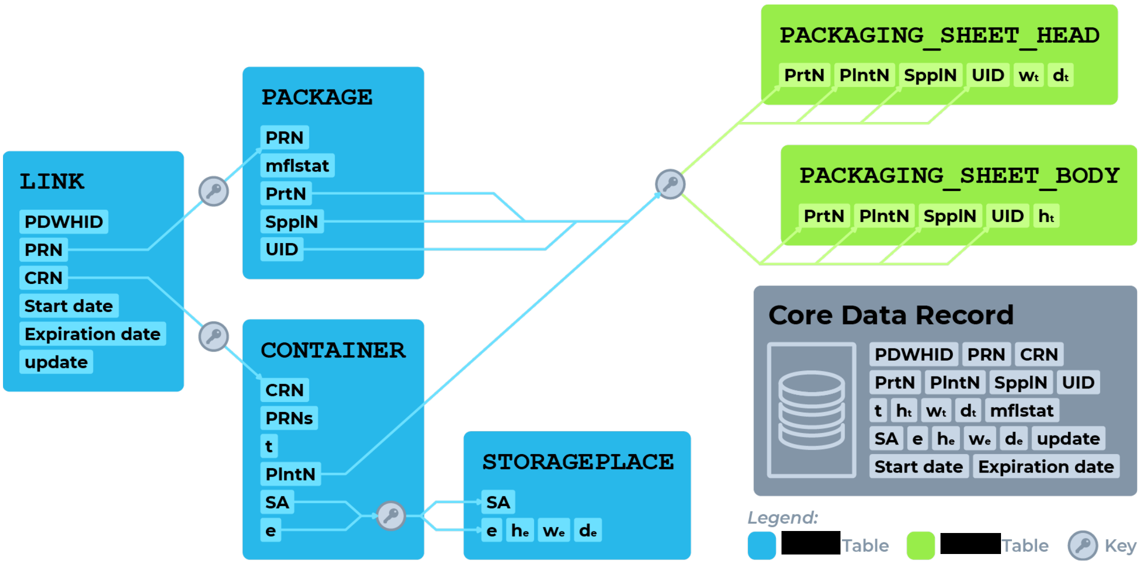


FIGURE 13. Data origins and connections

All these consequences naturally increase the number of evaluation possibilities for individual containers or packages. For example, a holistic history of the storage places of a CRN can be displayed using the data from the SL. At the same time, the SL also contains significantly more data records that have to be sorted, grouped, and filtered as part of the evaluation. On average, a PDC in Hanover North only contains around [redacted] actively stored containers and thus CRN in the ODS. However, in the SL, there are 5.4 million PDWHID for this particular PDC alone (in the whole data lake, there are 577.4 million containing approx. [redacted]). The second sprint of this project deals with the organization and filtering of this amount of data.

3.2.4 Dataset creation

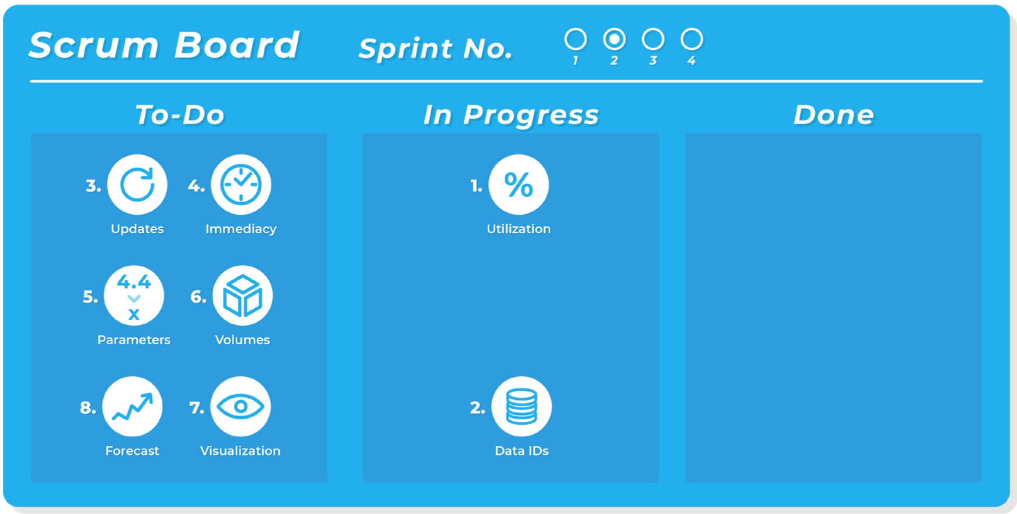


FIGURE 14. Scrum board at sprint 2

In order to correctly evaluate the PDC utilization in QuickSight, the target databases now had to be addressed after all required data had been identified in the LDL. Therefore, the Scrum board (Figure 14) nearly stays the same. Due to the large available base of data (63 different fields) in `LINK_SL` and the time-differentiated validity of the PRN-CRN connections, `LINK_SL` forms the backbone of the entire query. The goal of this sprint was to filter the 577.4 million data records to the approx. [redacted] containers currently stored in the PDC. In order to achieve this, the core dataset first had to be extended with data from the other tables so that exclusion criteria could then be applied. Of particular relevance in this step was the previous SAP Business Objects evaluation. Although this evaluation has some problems (described in Chapter 3.2.2), it also picks up all actually stored packages due to the exclusive access to `PACKAGE_ODS`. Thus, the data from `PACKAGE_ODS` serves as a reference and comparison point to check the data queried in the LDL for correctness and completeness. That was realized via a list comparison of the PRNs existing in the respective dataset. Therefore, this is iterative testing. In the following, challenges that arose in the course of this sprint will be categorized and highlighted as examples while at the same time presenting the appropriate approaches to solving them.

Storage zones in the PDC not to be evaluated

The key location filter for the query is, of course, the PDC itself. The LDL contains all packages of all warehouses of Volkswagen AG worldwide. In order to identify a PDC precisely, the plant number PlntN is required first. Within a plant, there are different storage areas SA with the respective storage elements *e*. The PDC to be analyzed contains exactly one SA with the number '[redacted]'. However, it is not the entire storage area '[redacted]' that is of interest because it contains not only the racks, aisles, and block storages but also storage zones classified differently and not used for direct storage.

One of the four largest areas is the storage distribution zone. All the packages that have left the goods received area, i.e., already marked with *mflstat* '4', are temporarily stored but still need to be distributed to the storage elements. After a truck delivery, for example, this zone is particularly busy. Accordingly, it is not affected by the actual utilization of the PDC. Secondly, there is a separate storage zone for additional quality inspection. Individual containers of a batch delivery are brought into this zone as random samples or on suspicion to be rechecked for quality. Third, and especially large, is the relocation zone. As mentioned before, block storage areas, in particular, have to be restacked or relocated in part because the parts currently required for production may be located further back inside them. Again, this is not an official storage area to be included in the evaluation, although it is directly influenced by the general utilization of the warehouse itself. If the block storages are too full, more stock transfer operations are needed, and the relocation zone is also more heavily utilized. Finally, there is the counterpart to the storage distribution zone, the delivery distribution zone. Here, the packages requested in production are removed out of storage and prepared for outbound transport via the bridge transport route. All packages that are in one of these four zones at a certain point in time must not be included in the calculation since their utilization in the actual sense is only the operational result of the utilization of the original storage elements in the PDC. In the query development, these zones are filtered out via excluding `WHERE` conditions that affect *e* (Appendix 3, lines 496, 503).

Packages removed, deleted, and relocated

Another decisive factor for improving the accuracy of the dataset is the consideration of the material flow status. Although the PDWHID can be used to filter for PRNs with the *mflstat* '4', this only means that this PRN was stored in a specific storage place at a specific time. However, the majority of the packages stored in the past have, of course, recently been taken out again and consumed in production. Therefore, another exclusion criterion was developed. In the table `PACKAGE_SL` there is a field that contains information about the deletion of a package, the field *deldate*. For the main query, therefore, two nested subqueries are created first, which generate lists of valid PDWHIDs, which in turn are linked to the next larger query using the SQL operators `WHERE ... IN ...` and `WHERE ... NOT IN ...` respectively. Following this logic, only PDWHIDs may be queried whose corresponding PRNs exist in `PACKAGE_SL` and have an empty *deldate* field (`WHERE deldate IS NULL`). However, because deletion dates may also exist in `LINK_SL`, no PDWHIDs may also be retrieved that have an existing deletion date (`WHERE deldate IS NOT NULL`) in `LINK_SL`. Only after these two lists are combined, *mflstat* '4' is checked (Appendix 3, lines 504–518).

Nevertheless, there can still be multiple instances of the same PDWHID with *mflstat* '4' and all the criteria mentioned so far. The reasons for this are the relocation processes that have already been mentioned. If a container is removed from a storage place but not deported to production, no deletion date is created. When moving to another storage place, instead a second data record of this CRN with a modified storage element *e* entry is generated. That must also be considered when developing the query. In this case, the *update* field finally comes into play. As part of the filtering of the dataset, a sub-list is created again, partitioned by PDWHID. In a group of equal PDWHID, ranks are assigned, starting from '1', ascending by descending update dates. That means that always the newest data record of a PDWHID, i.e., the one after the last relocation and thus with the current storage place, receives the rank '1'. Subsequently, only the PDWHIDs that have been assigned a rank of '1' can be allowed from this sub-list for further queries (Appendix 3, lines 494, 525).

Correlation and congruence check, time disparities

During the last iterations of the query development for the [redacted] tables with approx. [redacted] data records per query, more progress was achieved with the comparison data from `PACKAGE_ODS`. The PRNs were checked for congruence every 24 hours. Non-congruent PRNs, i.e., either PRNs from the Athena query that were not included in the original list (case 1) or PRNs from the original list that were not found via the Athena query (case 2), had to be examined more closely. First, correlation tests were performed on the PDWHID, the storage elements, and the PrtN. After this test, no interrelationship could be found between non-congruent PRNs and their storage places or contained part numbers. Unintentionally not recorded storage zones or obsolete parts could thus be excluded.

However, the non-congruent PRNs showed a tendency to lower PDWHIDs. PDWHIDs are assigned in ascending order according to their creation, so a temporal relationship was likely. In the following test loop, the correlation to the validation dates, the *update,* and the storage time was checked accordingly. While the storage date showed only a weak correlation to the non-congruence, a direct dependence of the other two fields could be recorded as a case distinction. Case 1 could be resolved via the validation dates. 10% of the excess data records were PDWHIDs whose connection validity had already expired, i.e., whose expiration date is in the past (Appendix 3, line 500). Case 2 showed that the closer the update of a PRN is to the actual query from `PACKAGE_ODS`, the more likely it belongs to the congruent PRNs. Discrepancies arise here due to the different update frequencies from the ODS and the SL. While ODS represents real-time storage, the data records in the SL are updated only once a day. If the observation and comparison period is extended to the following working day, all case-2 PRNs are now also congruent with the original ODS query. That completes the correct recording of all PDWHIDs.

Interference by PRNs of other plants

During the dataset creation process, there was a sudden increase in the number of queried data records after the table `PACKAGE_SL` was linked to `LINK_SL` via the PRN. The problem is that PRNs are historically not assigned unambiguously in the Volkswagen Group. The `LEFT JOIN` pulls every available match from the specified secondary table. If multiple matches exist, the left data record – in this case from `LINK_SL` – is simply picked up multiple times and joined with the different data records from the secondary table. This is exactly what happened, so that information about packages from all other Volkswagen AG locations was unintentionally added to the core dataset. However, this can be easily circumvented by adding the condition of a matching PlntN to the `LEFT JOIN` (Appendix 3, lines 520–522). This way, the dataset is definite again.

Type conversion and cleanup for [redacted] join

Unfortunately, the [redacted] information system knows neither PRNs nor CRNs. As mentioned in the previous chapter, the use of a key consisting of four fields (PrtN, SpplN, PlntN, UID) is thus necessary to connect `PACKAGING_SHEET_HEAD_SL` and `PACKAGING_SHEET_BODY_SL` to `LINK_SL`. However, these fields' data types and character lengths are not necessarily identical in [redacted] and [redacted] The numbers are mostly stored as strings and not as integers. That results in some fields not being identified as a match directly. A single excess blank space already leads to a mismatch. Therefore, a cleanup is necessary for some fields using the `trim()` function, which automatically trims spaces at the beginning and end of a string (Appendix 3, lines 471–473, 478). Another adjustment that had to be made involves truncating the [redacted] SpplN. The SpplN from [redacted] is supplemented by an additional code number at the end, which contains more detailed information about the contract with the supplier. That is a result of supply chain management, against the background of which only first-tier suppliers deliver directly to Volkswagen plants (Chapter 2.2). Nevertheless, individual components from second-tier suppliers are sometimes required without further processing in the value chain or are additionally distributed via a first-tier supplier. This code number shows whether such a component is an additionally traded part within the scope of a supplier contract. However, since this code number cannot be added automatically in [redacted] data records, it must be truncated in the data record from [redacted] (Appendix 3, line 479).

3.2.5 Data transfer and algorithm development

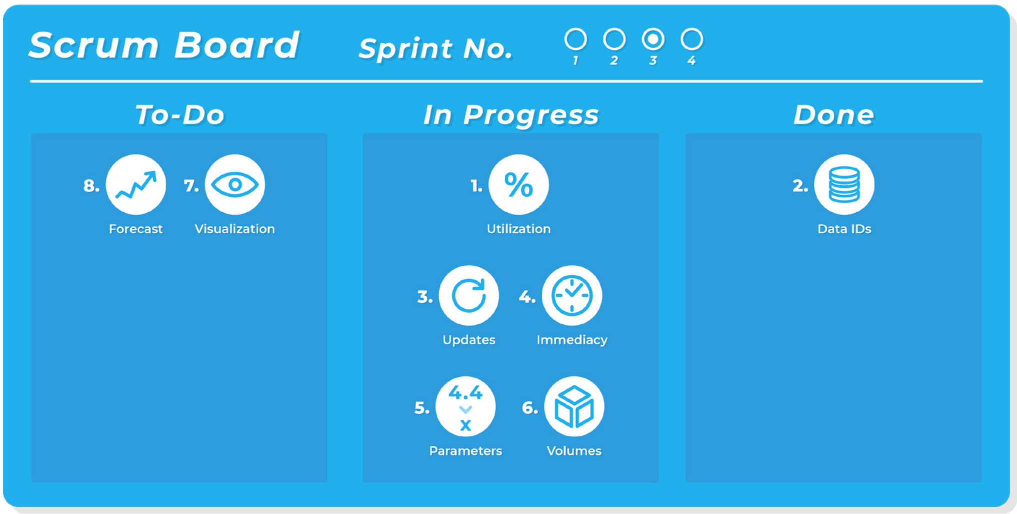


FIGURE 15. Scrum board at sprint 3

Now that the dataset has been created correctly and congruently with `PACKAGE_ODS`, the third sprint was to transfer the Athena dataset to Amazon QuickSight in order to ultimately derive the information required as a result (PBP 1) and also process the remaining points of the product backlog (Figure 15).

QuickSight supports a whole range of different data sources for generating datasets, including well-known names such as Salesforce, Oracle, MariaDB, Adobe Analytics, and, of course, Athena. In fact and among others, CSV (comma separated values) and XLSX files, Excel files based on the XML (expansible markup language) programming language developed especially for data exchange, are also allowed as data sources. The apparent disadvantage of the last two formats is that these files do not update themselves, let alone connect to a real database. Predictably, Athena had to be chosen as the source here in order to be able to integrate the developed queries (Appendix 3) targeting S3. After QuickSight has established the connection to Athena, all tables available for the data scientist role can now be selected directly. However, since no table can be used as a stand-alone object for the analysis, a custom Athena SQL query must be established to access the designed dataset.

Once an initial data source is available as a dataset, all sources and their relationships to each other can be managed in QuickSight. Each individual query to a database is considered a sub-dataset, which in turn can be connected via `JOIN` links. A crucial function can already be activated in this interface. The query mode can be switched to the already mentioned SPICE in-memory computing, which increases the query and evaluation speed for the reasons mentioned before. That plays into the hands of PBP 4, immediate availability. One of the most important reasons for choosing Amazon Web Services, especially QuickSight, as a solution to correct warehouse utilization analysis is the possibility to have all data sources of a dataset automatically updated at regular intervals. Therefore, the various queries are sent to the databases at times determined by the user without any further intervention needed on the part of the user. That successfully concludes PBP 3, automatic updates.

For the solution of PBP 5 and 6, the optimized volume-oriented utilization calculation, a holistic redesign of the calculation algorithm is necessary. For this purpose, a distinction is again made between rack, aisle, and block storage, whereby the aisle storages will form the basis for the calculation of both other storage elements. The principle for the optimized calculation of utilization κ is described: According to the container distribution *$P_e(T)$* of a particular storage element in relation to the total container distribution of the PDC *$P_{total}(T)$*, a theoretically still available volume in a storage element *$V_{e,theor}$* is calculated, which is then compared with the actually occupied volume *$V_{e,occ}$*. Thus, for the individual utilization of an element as well as for the total utilization of the PDC, the following applies:

$$\kappa_e = \frac{V_{e,occ}}{V_{e,theor} + V_{e,occ}} \tag{9}$$

$$\kappa = \frac{\sum_{e \in E} \left(V_{e,occ}\right)}{\sum_{e \in E} \left(V_{e,theor} + V_{e,occ}\right)} \tag{10}$$

Two premises apply to the following calculation steps, which are valid in the Hanover North PDC. First, as before and according to their stacking factor *s_t,* only containers of the same type can be stored on top of each other. And second, that all container types appearing in the warehouse are too large in their dimensions, i.e., both *w_t* and *d_t*, for any combination of two containers to fit next to each other in the 1.20 m wide aisles. Thus, they can only be stored lengthwise along the aisle length *l_e* or on top of each other under the condition mentioned above.

The first step is to determine the occupied length of an aisle $l_{e,occ}$. For this purpose, it must be counted how many stacks of each container type appearing in the storage element have been started ($n_{st,t,e}$). For this, simply the rounded up number from the division of the number of containers of a type t existing in the element e ($n_{t,e}$) with the associated stacking factor s_t is used (Formula 11). The occupied length $l_{e,occ}$ results accordingly from $n_{st,t,e}$, and the associated x_t, i.e., either w_t or d_t (Formula 12). For the determination of x_t, a case distinction must be made. If w_t and d_t are both smaller than the aisle width of 1.20 m, x_t is the minimum of both values. If one of the values is higher than 1.20 m, this value is automatically x_t.

$$n_{st,t,e} = \left\lceil \frac{n_{t,e}}{s_t} \right\rceil \tag{11}$$

$$l_{e,occ} = \sum\nolimits_{t \in T} (n_{st,t,e} \cdot x_t) \tag{12}$$

A further case distinction for calculating the theoretically available free volume of a storage element $V_{e,theor}$ follows. On the one hand, there are free volumes of unfinished stacks in which containers of the same type would still fit ($V_{e,1}$), and also a residual length free volume, i.e., the total volume on the unoccupied length of the aisle ($V_{e,free}$). $V_{e,1}$ can be calculated by multiplying the number of containers still missing to complete a given stack of a particular container type t ($n_{free,t,e,1}$) by the corresponding dimensions:

$$V_{e,1} = \sum_{t \in T} n_{free,t,e,1} \cdot h_t \cdot w_t \cdot d_t \tag{13}$$

However, the residual length free volume $V_{e,free}$ cannot be included in $V_{e,theor}$ directly, since it is very unlikely to be used completely. The densest packing of all containers is only conceivable in the hypothetical ideal case. Therefore, a theoretically usable volume $V_{e,2}$ is calculated here, which can be determined from the containers that could still be placed in the residual length free volume $V_{e,free}$. For this purpose, it must be decided which container types and how many entities of them could be placed in the volume. In order to get there, the container type distributions of a storage element $P_e(T)$ and the total distribution $P_{total}(T)$ are consulted. If an element is already mostly occupied (e.g., $l_{e,occ}$ is approx. 80% of l_e)

the storage should be continued with a similar distribution if possible. If, however, it is still relatively free, it should be stored rather according to the total PDC distribution $P_{total}(T)$. To realize this model, a theoretical interim distribution $P_{e,theor}(T)$ is needed, which is determined by the ratio of $l_{e,occ}$ to l_e:

$$P_{e,theor}(T) = \frac{l_{e,occ}}{l_e} \cdot P_e(T) + \frac{l_e - l_{e,occ}}{l_e} \cdot P_{total}(T) \tag{14}$$

Then, according to the new distribution $P_{e,theor}(T)$, it is ascertained how many containers of a type still fit into the free residual length free volume at maximum possible stacking in each case ($n_{free,t,e,2}$). The process starts with a stack of the most frequent container type t_1 from $P_{e,theor}(T)$, followed by the second most frequent type t_2 and so on, sticking to the distribution as close as possible. These process steps are repeated until the free length is filled with the hypothetical stacks and their corresponding x_t. This results in Formulas 15 and 16:

$$n_{free,t,e,2} = \left\lfloor \frac{(l_e - l_{e,occ}) \cdot P_{e,theor}(T = t)}{x_t} \right\rfloor \cdot s_t \tag{15}$$

as long as

$$l_e - l_{e,occ} \geq \sum\nolimits_{t \in T} \frac{n_{free,t,e,2}}{s_t} \cdot x_t \tag{16}$$

is true.

After these steps have been performed successfully, the individual utilization κ_e and the total utilization κ can be finally calculated using the composite theoretical volume $V_{e,theor}$ (Formula 17) and the actual occupied volume $V_{e,occ}$ (Formula 18) with the Formulas 9 and 10 from the beginning. A visual representation of $V_{e,1}$ and $V_{e,2}$ shows Figure 16.

$$V_{e,theor} = V_{e,1} + V_{e,2} = \sum\nolimits_{t \in T} (n_{free,t,e,1} + n_{free,t,e,2}) \cdot h_t \cdot w_t \cdot d_t \tag{17}$$

$$V_{e,occ} = \sum_{t \in T} n_{t,e} \cdot h_t \cdot w_t \cdot d_t \tag{18}$$

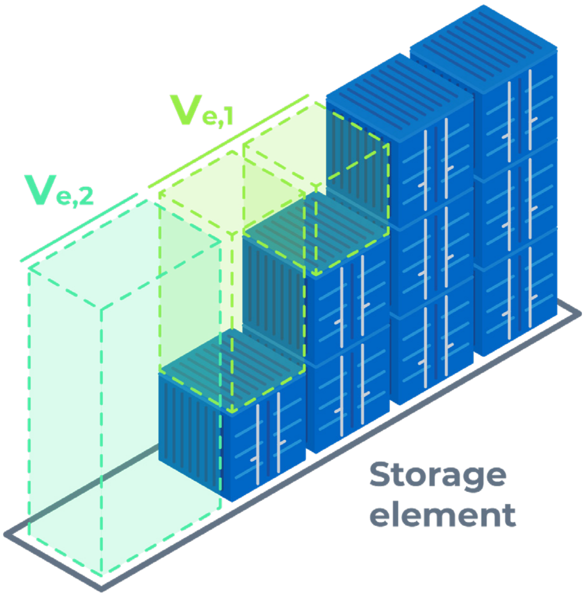


FIGURE 16. Theoretic volumes in a storage element

This is the algorithm for calculating the stock utilization based on the aisles. Since block storages make up only a small part of the PDC (<10% of the storage area), a simplification is made. The length of the shorter one of each sides of a block storage is divided by 1.20 m. Thus, the block storage is approximated as an aisle storage with l_e (Formula 19). Hence all following calculation steps are identical. This simplification can be made because only containers stacked by type are found even in the block storages.

$$l_e = \frac{\min{(w_e,d_e)}}{1.20\text{ m}} \cdot \max{(w_e, d_e)} \tag{19}$$

The calculation of κ_e for a rack element is very similar. However, the difference is that it is not the occupied length $l_{e,occ}$ compared to the total length l_e that matters, but the number of occupied rack compartments $n_{rc,occ}$ compared to the total number of all compartments available per rack, n_{rc}. That changes the calculation of $V_{e,1}$, since each rack compartment *rc* out of the set of occupied *rc* (RC_{occ}) must be considered individually. A premise made for this calculation is that only containers of one type *t* get stored in a single rack compartment. The number of these is named $n_{t,rc}$ (Formula 20). Also, the number of containers that can be stacked on top of each other is limited by the height h_{rc} of a rack compartment while the stacking factor s_t must not be exceeded (Formula 21).

$$V_{e,1} = \sum\nolimits_{rc \in RC_{occ}} \left(\left\lfloor \frac{h_{rc}}{h_t} \right\rfloor \cdot \left\lfloor \frac{w_{rc}}{x_t} \right\rfloor - n_{t,rc} \right) \cdot h_t \cdot w_t \cdot d_t \tag{20}$$

while

$$\left\lfloor \frac{h_{rc}}{h_t} \right\rfloor \leq s_t \tag{21}$$

The equations 14–16 must be adjusted as follows:

$$P_{e,theor}(T) = \frac{n_{rc,occ}}{n_{rc}} \cdot P_e(T) + \frac{n_{rc} - n_{rc,occ}}{n_{rc}} \cdot P_{total}(T) \tag{22}$$

$$n_{free,t,e,2} = \lfloor (n_{rc} - n_{rc,occ}) \cdot P_{e,theor}(T = t) \rfloor \cdot \left\lfloor \frac{h_{rc}}{h_t} \right\rfloor \cdot \left\lfloor \frac{w_{rc}}{x_t} \right\rfloor \tag{23}$$

while

$$n_{rc} - n_{rc,occ} \geq \sum\nolimits_{t \in T} \frac{n_{free,t,e,2}}{\left\lfloor \frac{h_{rc}}{h_t} \right\rfloor \cdot \left\lfloor \frac{w_{rc}}{x_t} \right\rfloor} \tag{24}$$

and Equation 21 are true.

With these values, κ_e for a rack and κ via the Formulas 17, 18, 9, and 10 can be calculated. Through this algorithm (Formulas 9–24), PBPs 5 and 6 are successfully fulfilled. The entire calculation is thus fully parameterized, based on datasets updated at regular intervals, and depend on the actual volume instead of the number of units and standard container sizes. Hence, PBP 1 has been achieved, too. The next step consequently deals with PBP 7, the visual processing of the results.

3.2.6 Information design

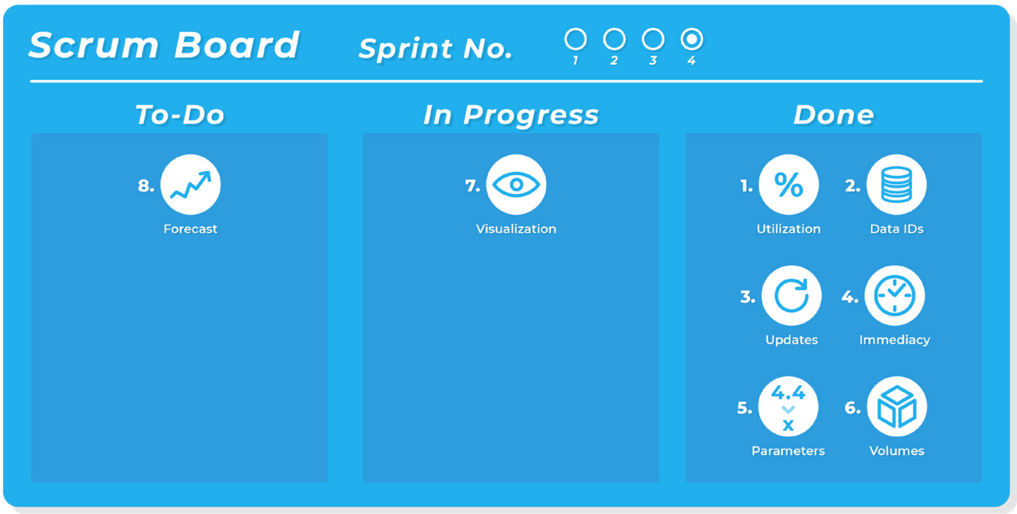


FIGURE 17. Scrum board at sprint 4

This project's fourth and final sprint focuses on visualization and interaction options with the results and information obtained from the data (Figure 17). The presentation in QuickSight was realized according to the top-down principle. That means that the most generalized information is shown first and then gradually goes into more detail. In concrete terms, this means that in the overview tab of the dashboard, the KPI warehouse utilization (PBP 1.a, Figure 18, 1) is displayed centrally as the very first thing, followed directly by the most heavily utilized storage elements in descending order (PBP 1.b, Figure 18, 2). After that, the current total container distribution is displayed (PBP 1.c, Figure 18, 3).

As in the algorithm itself, a distinction is made according to the type of storage element (aisle, block, rack), with deeper insights in three corresponding tabs (Figure 18, 4). In each of these, the individual storage elements and their occupation and utilization are presented. In addition, the distributions $P_e(T)$ and $P_{e,theor}(T)$ can be found here. The individual interim data, such as $n_{t,e}$ or $V_{e,theor}$ remain in the background for the time being. However, they can be added as an additional field via the menu bar (Figure 18, 5) if required. The generated QuickSight interface is shown in Figure 18.

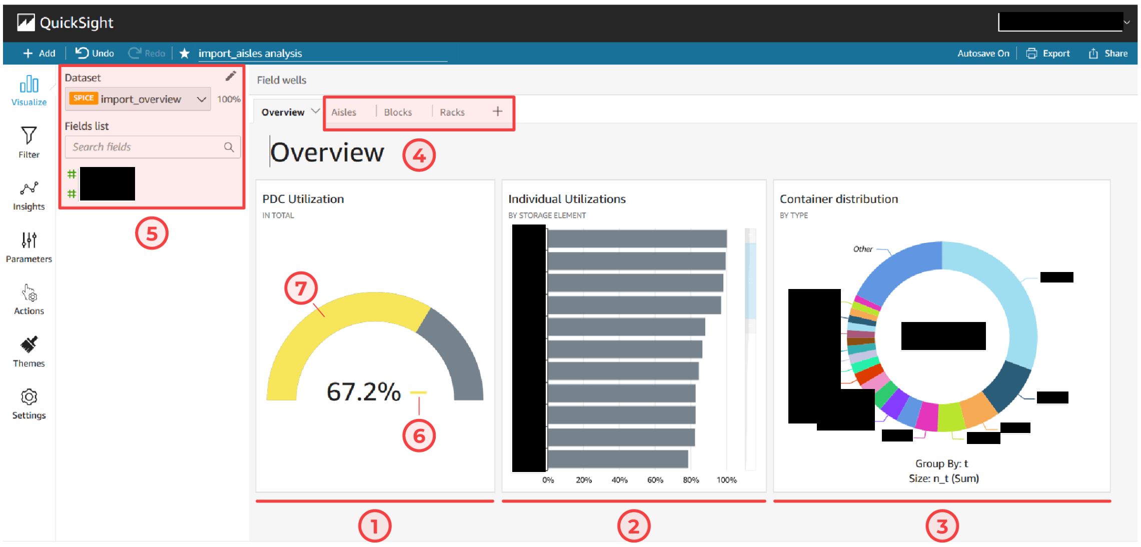


FIGURE 18. QuickSight overview tab

Problems with UX design in QuickSight

Amazon QuickSight offers a range of already mentioned advantages. As a solution within AWS and the possibility of directly linking to the data of the LDL on S3, it currently represents the best possibility for data analysis of the PDC. However, concerning this project and PBP 7, there are not to be dismissed deficits regarding the presentation of the obtained information. For example, conditional formatting in QuickSight is only available for certain visualization objects such as the individual KPI or the table display. These options are indeed beneficial. For instance, the visual layout depending on the overall utilization is practical and facilitates the quick creation of an overview. It automatically generates trend icons (Figure 18, 6) and adaptive color selection (Figure 18, 7) when utilization is too high or too low (fulfilling UX points 2 and 3, ease of orientation and efficiency).

However, the implementation as a whole does not correspond to the original ideas of the graphical preparation of the obtained information. The central element was supposed to be an approximated workshop layout of the PDC, showing the different storage zones and storage elements in scale and reflecting their individual workloads. Furthermore, the idea was to realize intuitive interactivity according to UX principles. For example, by selecting individual elements in the PDC layout, it should be possible to navigate to the desired sub-information such as $P_e(T)$ and $P_{e,theor}(T)$ of an individual storage element while also providing the core data on demand.

Alternative solution approach

As the last step of the fourth sprint, a concept for an improved visualization was created for these reasons. The basis of this concept is still datasets generated using SQL. However, the entire implementation is not done via a PaaS tool such as QuickSight, but rather an especially newly designed website using HTML (hypertext Markup Language), PHP (hypertext preprocessor), JavaScript, and SQL. The design is consequently done using CSS (cascading style sheets). The complete concept, including basic functionalities, can be found in Appendix 4. Figure 19 shows the overview menu, which already explains the advantages of UX design compared to QuickSight.

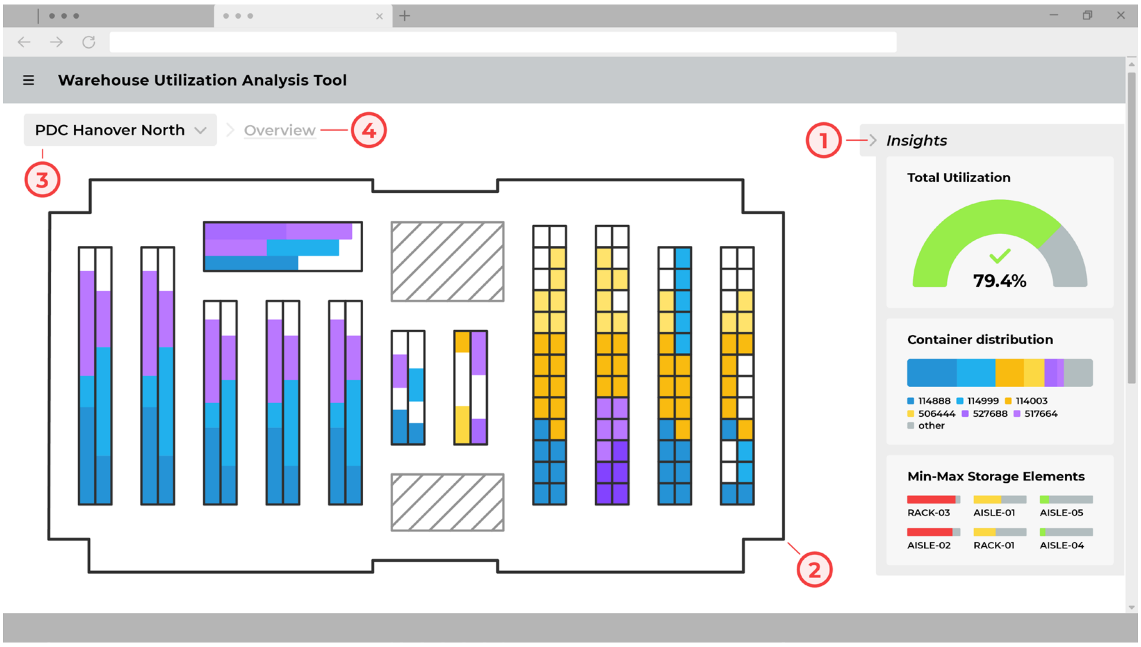


FIGURE 19. Design concept for the warehouse utilization analysis tool

The first meaningful change is reducing the number of buttons for the user directly on the first page. That improves learnability (UX point 1) and ease of orientation (UX point 2). According to lean UX design, only the absolutely crucial functionalities demanded by the end user are implemented (PBP 1–7). The insight section can be collapsed (Figure 19, 1), if necessary, to view the layout (Figure 19, 2) in large, which is the primary display element of the tool. The layout is intentionally presented in an abstract way. For example, not all walls, entrances, forklift or escape routes are shown. In this project, only the information about the individual storage elements is decisive.

A warehouse to be evaluated, such as the PDC Hanover North in this case, is selected via the drop-down list (Figure 19, 3) right at the beginning of the page. If the user is interested in the utilization, occupancy, or container distribution of a particular element, this can simply be selected directly in the layout (Appendix 4). That saves the time-consuming search in tables and lists for the sometimes quite cryptic designations for racks and aisles (Figure 18, 2). In turn, this increases efficiency (UX point 3), while the spatial segmentation of information also improves memorability (UX point 4). In order to ensure accessibility (UX point 5) with this high image-to-text ratio, there is still a list of storage elements that screen readers can read out. However, the difference to QuickSight is that corresponding pairs, i.e., list entry and storage element in the graphic, are always highlighted when hovering over them.

Incorrect entries and accidental "moving" of records or insights are not possible for an end user, as these elements are firmly integrated into the page. However, suppose they select a storage element for detailed analysis that they did not really want to inspect. In that case, they can quickly return to the 'overview' page via the navigation bar above (UX point 6, error forgiveness, Figure 19, 4). Finally, the use of the application should also bring delight (UX point 7), which is realized by so-called micro-animations and micro-interactions. That means all transitions, such as the folding or unfolding of insights or the transfer to a single storage element, run smoothly and not abruptly.

Technically, these functionalities are mainly realized via an adaptive SVG (Scalable Vector Graphic) embedded in the HTML code, i.e., the graphic type of the warehouse layout controlled and updated via JavaScript. The individual elements of the SVG and also the storage element list entries are assigned HTML data attributes corresponding to their storage element names. These are checked on interaction, e.g., a mouse hover or click event. A database connection is to be made with the help of PHP. However, this is also the current main hurdle for the realization of this concept. The S3 data of the Logistics Data Lake can currently only be accessed via the AWS tools (provider exclusivity). In addition, there is, of course, no certification of a Volkswagen PKI Certification Authority yet and accordingly no public key for such an application. Therefore, all representations of this concept are only filled with dummy data. The use of internal data is prohibited at this stage of development. Nevertheless, implementing such a concept or at least a similar one is recommended in the long term.

3.3 Forecast development and outlook

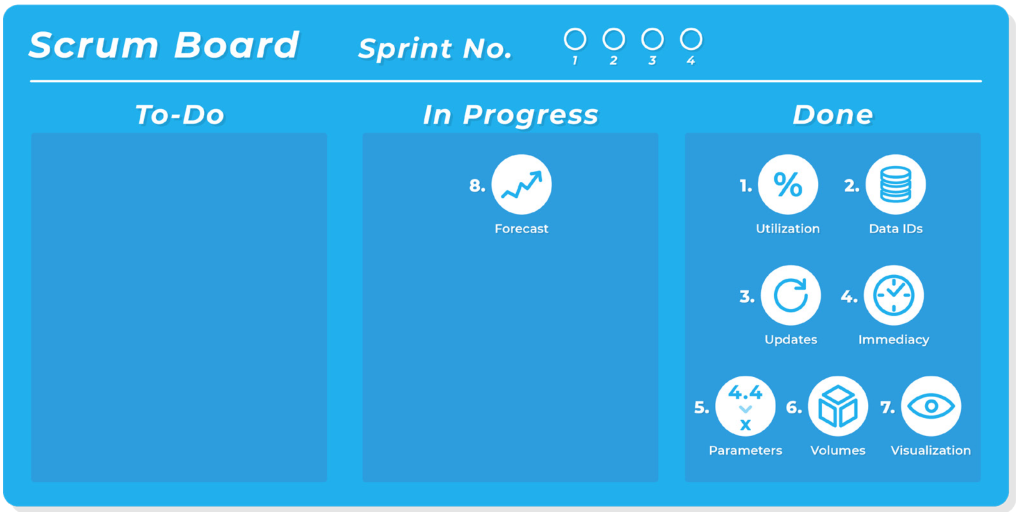


FIGURE 20. Scrum board after sprint 4

After every PBP (except from the low-priority PBP 8) had been processed and the last one of the four sprints had been run, the project was considered officially completed (Figure 20). Nevertheless, the request for a forecast remains. After the four sprints, thoughts were given to this topic, and possible solutions were drafted. There are two conceivable variants for the realization of the forecast.

The first variant is to use QuickSight machine learning insights. An advantage here is that no own algorithms must be designed. Against the PaaS/SaaS background, the corresponding tools can be applied to entire datasets via drag and drop (Amazon Web Services n.d.b). However, an essential premise for this is that sufficient data for pattern recognition is available, which is not the case for this project yet. Again, there are two ways to change this. The simpler version is to wait for matching future evaluations by continuously recording single values like the individual utilization of the storage elements. The more complex (and desirable) version would be to segment the existing historical data from the LDL by individual days and process it in such a way that it becomes usable for QuickSight machine learning insights. That requires an appropriate segmentation strategy.

A looped SQL query that combs through all past data based on the update field or the storage time field would be possible here. Segmentation by the `PARTITION BY` function within a single SQL query in combination according to the two previously mentioned fields is a recommended option as well.

The second and more precise variant, especially for the time horizon of a few days, is implementing the planned deliveries of the suppliers. Ideally, this would require the real-time data of all suppliers to be made available as planned in the group strategy regarding the Volkswagen Industrial Cloud. In this way, it would also be possible to determine in advance from the LDL data how many containers of which type will arrive on the next day. In line with supply chain management, the stocks in the PDC could also be reported back to the suppliers in real-time. Hence, the material flow can be controlled in a feedback loop, and over-utilization or under-utilization can be avoided much earlier in the value chain.

However, even in this case, the application of machine learning tools would be worthwhile because the planned delivery quantity is only the optimal number of containers. Of course, goods to be returned due to quality defects are not considered here. In addition, delivery on a particular day does not necessarily mean that the container will be taken to a fixed storage location on the same date. It also happens that a container first remains in one of the storage zones mentioned in Chapter 3.2.4, which are not relevant for the calculation. Therefore, in this context, either empirical values would have to be used, or a prediction based on historical data would have to be made. Of course, it would also be particularly interesting to know whether there are any differences between individual container types in this respect. In turn, this information could serve as a basis for further optimization, e.g., through container adjustments.

These preparations for forecast development lead to further considerations regarding future applicable technologies for warehouse analyses. One of the most apparent optimizations is installing RFID transponders instead of barcodes for container identification. Since all containers are integrated into a return cycle, no transponders are lost (emptied containers are transported back to the supplier by the warehouse's service provider). RFID offers some advantages compared to barcodes. Of particular relevance for this project would be that the chips can be read without visual contact and can store information themselves (Logistik KNOWHOW 2017). While the barcode methodology from Figure 9 creates a new data record with nearly the same information on each zone change and relocation process, an RFID chip could store the material flow statuses that have already passed through. Sprint number two of this project could have been completed much faster with this technology. The identification process is also more straightforward because no optical recognition is necessary but instead takes place via radio frequency reception. Therefore, in some cases, scanning processes could be accelerated by a factor of 20 because multiple RFID chips can also be detected simultaneously by a single receiver (Logistik KNOWHOW 2017). Of course, as a more far-reaching outlook, the investment costs for converting the containers to RFID must be considered and offset against potential savings in terms of storage efficiency as a next step. A further outlook regarding the evaluation criteria is presented in the following chapter.

4 CONCLUSION

4.1 Project result evaluation

After the project has been completed, it is necessary to assess the extent to which it has been successful and whether the results can be applied to other areas. The results were assessed based on four different criteria. These are the completed product backlog points, the data accuracy, the applicability to other purchasing and distribution centers, and the utilization accuracy.

Product Backlog Points

After completing the project, the first seven of the eight PBPs can be assigned the status 'Done'. This means that the majority of the tool requirements prioritized by the product owner have been implemented successfully. The utilization of the PDC is analyzed in total (PBP 1.a), whereby individual values for the respective storage elements are also displayed (PBP 1.b). The various container distributions are also determined (PBP 1.c). The data required for this were identified in the Logistics Data Lake, and suiting datasets were formed from them (PBP 2). These datasets update automatically in QuickSight (PBP 3) and are immediately available via the share function of the interactive dashboard for all employees who have access via PKI Card (PBP 4). The calculation is no longer based on the assumption of static values but on parameters (PBP 5).

However, there is a catch to be mentioned here. The dimensions of certain storage elements are not wholly entered in the LDL since the system is still being built up in the Volkswagen Industrial Cloud. Therefore, these parameters are not yet pulled directly from the LDL but are still obtained from an interim spreadsheet uploaded in QuickSight. In order to gain access to the fully parameterized data, a system interface to the service providers [redacted] and [redacted] will be necessary in the long term. The service providers, as operational forces in the warehouse, adjust these dimensions independently. This interface was still being planned at the time the project ended.

Nevertheless, the calculation is now volume-based and not unit-based (PBP 6). A visualization concept in QuickSight was implemented (PBP 7). Another optimized visualization concept as a stand-alone web application was also designed but is not officially used yet. The reason for this is a missing PKI certification. Preparations for a forecast of future utilizations have been made (PBP 8). This point should be implemented within the next two months.

Data accuracy

The term data accuracy refers to the accuracy of the data basis. More precisely, this means the percentage of containers actually stored in the PDC that appear in the data records extracted from the LDL. Neither missing containers nor surplus containers that do not count for the evaluation may be recorded. According to this measurement, the data accuracy is 100%. This number is determined by comparing the queried data with the data from the operational data store, the real-time system in which only data of currently available containers are stored. During an iterative testing process, the individual data sets of different days were compared with each other until only 100% matches existed. However, there is a disadvantage of the LDL storage layer compared to ODS. The data is updated only once every 24 hours. This means that time disparities may occur. In extreme cases, a container that has been in storage for 23 hours and 59 minutes could not be included in the evaluation. But since the ODS evaluation was previously only compared with production planning once a day, this is not a currently prioritized problem. However, in the future, it is planned to include the ODS in the LDL so that the analysis can also be performed in real-time if required.

Applicability for other warehouses

This evaluation criterion is completely satisfied. In the SQL code, only the two numbers identifying the storage areas must be adjusted to access data from another PDC (Appendix 3, lines 498, 510). This has already been tested successfully with the PDC Hanover West. Since the data structure is mostly the same, the identification criteria determined in Sprint 1 and 2 of this project are also applying. A simple evaluation of other warehouses is thus guaranteed. However, there is still a need for differentiation here. If storage elements other than aisle, rack, and block storages appear, the calculation algorithm developed in Sprint 3 is no longer fully effective. A more extensive calculation, e.g., for automatic miniload systems, still has to be developed.

Utilization accuracy

The request for an optimized warehouse utilization data analysis originally came from the responsible warehouse supervisors at Volkswagen AG. The problems of the previous calculation described in Chapter 3.2.2 had already been suspected at that time. A resulting problem was the question of responsibility between Volkswagen AG and the service providers of the warehouse. If, for example, the warehouse utilization is 90%, which corresponds to an over-utilization, the service provider makes demands for additional payments. Because the warehouse is rigid, more relocation processes are necessary, which requires additional shifts or staff. According to the impression of the warehouse supervisors, however, these demands were made too quickly in the past. In other words, this means the percentage calculated so far seemed too high in comparison to the observed warehouse occupation. According to initial results, the calculation method developed within this project's scope indeed shows a trend towards an approximately 5% lower utilization. However, this trend has still to be monitored and officially acknowledged by both sides, i.e., Volkswagen and the service providers. When this is finally implemented depends heavily on the respective service provider. The derivation of the adjusted calculation explained in this thesis paper shall support this process.

4.2 Summary

The trends of the fourth industrial revolution, Industry 4.0, are significantly shaping the development of the automotive industry. Cyber-physical systems networked by the Internet of Things generate exponentially growing amounts of data that must be managed appropriately. A proven solution approach for companies not necessarily specialized in information technology is cloud computing. That means renting online storage and processor capacity from suitable providers tailored to the actual requirements of an individual user.

The cooperation between Volkswagen AG and Amazon Web Services (AWS) represents such a system. As part of the Volkswagen Industrial Cloud project launched in 2019, solutions for the utilization and analysis of big data are jointly developed based on the AWS platform (platform as a service). One concrete application of this is the Logistics Data Lake (LDL), in which not only all logistics databases of Volkswagen AG will be combined in the long term, but also the data of all suppliers delivering to Volkswagen will be implemented.

This large-scale collection of data offers completely new possibilities to derive information from it. Accordingly, the objective of this bachelor's thesis was to exploit one of these potentials. A warehouse utilization data analysis and forecast tool was developed within this project.

For this purpose, the required data had to be identified and accessed in the LDL first. Mainly the two logistics information systems [redacted] and [redacted] were accessed as data sources via SQL queries in Amazon Athena. The goal of these queries was to retrieve datasets with information about the containers stored in a warehouse and the installed storage elements. After the corresponding identification was developed as a combination of several IDs, duplicates and unnecessary data records had to be filtered out of the datasets. In order to achieve this, a real-time operational data store was used as a reference. In an iterative test procedure, additional exclusion criteria were developed until the datasets were suitable. The datasets were then transferred to Amazon QuickSight and evaluated using a specially developed algorithm. The final step focused on visualization for comprehension by a human user since the obtained information is relatively complex.

The agile project management method Scrum was used for the entire project. For this purpose, defined roles were assigned in the project team at the beginning, and a product backlog was created. It contained demanded functionalities that the tool should have. Within four three-week sprints, the items of this product backlog were processed according to their prioritization. The entire process of information derivation was carried out using the intelligence cycle within the structured data analysis.

A 100% data accuracy respecting the stored stock of a sample warehouse in Hanover could be achieved while other warehouses could also be successfully analyzed afterward. However, the accuracy of the utilization percentage still needs to be monitored and confirmed by the warehouse supervisors who have requested the new analysis tool.

In addition, preparations have been made for AI-based forecast development. Its implementation is still pending due to the current lack of a matching data basis. Nevertheless, it is to be realized in the near future. The project itself was thus rated as a success by the product owner.

REFERENCES

Amazon Web Services n.d.a. Amazon Athena. Start querying data instantly. Get results in seconds. Read on 05.05.2021. https://aws.amazon.com/athena/?nc1=h_ls&whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc.

Amazon Web Services n.d.b. Amazon QuickSight ML Insights. Read on 05.05.2021. https://aws.amazon.com/de/quicksight/features-ml/.

Amazon Web Services n.d.c. Amazon QuickSight. Scalable, serverless, embeddable, ML-powered BI Service built for the cloud. Read on 05.05.2021. https://aws.amazon.com/quicksight/?nc1=h_ls.

Amazon Web Services n.d.d. Amazon S3. Object storage built to store and retrieve any amount of data from anywhere. Read on 29.04.2021. https://aws.amazon.com/s3/?nc1=h_ls.

Amazon Web Services n.d.e. Was ist Cloud Computing? Read on 17.04.2021. https://aws.amazon.com/de/what-is-cloud-computing/.

Cleff, T. 2011. Descriptive statistics and modern data analysis. A computer-based introduction with Excel, PASW (SPSS) and STATA. 2nd ed. Wiesbaden: Springer Gabler.

Fleig, J. 2019. Agiles Projektmanagement. business-wissen.de. Read on 06.05.2021. https://www.business-wissen.de/artikel/agiles-projektmanagement-so-funktioniert-scrum/.

Gronbach 2018. Statusanalyse: Der Systemlieferant innerhalb der Zuliefererpyramide. Gronbach Magazin. Read on 07.04.2020. https://www.gronbach.com/blog/business/statusanalyse-der-systemlieferant-innerhalb-der-zulieferpyramide?noredirect=de_DE.

Haak, S. 2019. VW will seine Fabriken mit Amazon-Technologie ausstatten. Business Insider Deutschland GmbH. Read on 12.04.2021. https://www.businessinsider.de/gruenderszene/automotive-mobility/vw-amazon-partnerschaft/.

Hennecke, F. 2015. Lean UX: Mit Start-up-Methoden zu einem besseren Produkt. Heise Medien. Read on 29.04.2021. https://m.heise.de/developer/artikel/Lean-UX-Mit-Start-up-Methoden-zu-einem-besseren-Produkt-2544807.html?seite=all.

Hermann, M., Pentek, T. & Otto, B. Design Principles for Industrie 4.0 Scenarios. In: 49th Hawaii International Conference on System Sciences (HICSS). 2016, 3928–3937. doi: 10.1109/HICSS.2016.488.

Hinterberger, H. Exploratory Data Analysis. In: L. Liu/M. T. Özsu (Eds.). Encyclopedia of Database Systems. 2009. Boston, Springer. https://doi.org/10.1007/978-0-387-39940-9.

Hummel, T. 2019. Praxishandbuch JIT/JIS mit SAP®. Die Just-in-Time und Just-in-Sequence Abwicklung mit SAP®. Berlin: Springer Vieweg.

Hundertmark, H. 2013. Beziehungsmanagement in der Automobilindustrie. OEM Relationship Management als Sonderfall des CRM. Wiesbaden: Springer Gabler.

Hüttel, Heiko 2019. Volkswagen Automotive Cloud and Microsoft. Interviewed by Volkswagen AG 2019. Read on 12.04.2021. https://www.volkswagenag.com/de/news/stories/2019/03/automotive-cloud-volkswagen-and-microsoft-develop-mobility-ecosy.html.

Jungwirth, K. 2016. Agile Methoden: So funktioniert Scrum. InLoox. Read on 07.04.2021. https://www.inloox.de/unternehmen/blog/artikel/agile-methoden-so-funktioniert-scrum/.

Kamps, U. 2018. Datenanalyse. In: Gabler Wirtschaftslexikon. n.d. Read on 29.04.2021. https://wirtschaftslexikon.gabler.de/definition/datenanalyse-30331/version-253916.

Logistik KNOWHOW 2017. Advantages of RFID technology over barcodes. Dr. Thomas + Partner GmbH & Co. KG. Read on 04.05.2021. https://logistikknowhow.com/bestandsverwaltung/rfid-vorteile-gegenuber-dem-barcode/.

Luber, S. 2016. Was ist das Internet of Things? BigData Insider. Read on 17.04.2021. https://www.bigdata-insider.de/was-ist-das-internet-of-things-a-590806/.

Luber, S. 2017. Was ist ein Cyber-physisches System (CPS)? BigData Insider. Read on 17.04.2021. https://www.bigdata-insider.de/was-ist-ein-cyber-physisches-system-cps-a-668494/.

Lynch, P. J. & Horton, S. 2016. Web Style Guide. Foundations of User Experience Design. 4th ed. New Haven: Yale University Press. Read on 07.04.2021. https://webstyleguide.com/

Microsoft Corporation 2021. Volkswagen Konzern und Microsoft beschleunigen Entwicklung des automatisierten Fahrens. Read on 12.04.2021. https://news.microsoft.com/de-de/volkswagen-microsoft-beschleunigen-entwicklung-des-automatisierten-fahrens/.

Oracle n.d.a. Database. Read on 06.04.2021. https://www.oracle.com/de/database/what-is-database.html.

Oracle n.d.b. Was ist Big Data? Definition von Big Data. Read on 17.04.2021. https://www.oracle.com/de/big-data/what-is-big-data/.

Platform Industrie 4.0 2021. The background to Plattform Industrie 4.0. Ziel, Struktur und Geschichte der Plattform. Federal Ministry for Economic Affairs and Energy/Federal Ministry of Education and Research. Read on 12.04.2021. https://www.plattform-i40.de/PI40/Navigation/EN/ThePlatform/Background/background.html.

Radtke, M. 2019. Was ist Big Data? BigData Insider. Read on 17.04.2021. https://www.bigdata-insider.de/was-ist-big-data-a-562440/.

Siepermann, M. & Lackes, R. 2018. Business Intelligence. In: Gabler Wirtschaftslexikon. n.d. Read on 10.05.2021. https://wirtschaftslexikon.gabler.de/definition/business-intelligence-29438/version-253044.

Statista 2021a. Größte Automobilhersteller weltweit nach Fahrzeugabsatz im Jahr 2020. Statista GmbH. Read on 12.04.2021. https://de.statista.com/statistik/daten/studie/173795/umfrage/automobilhersteller-nach-weltweitem-fahrzeugabsatz/.

Statista 2021b. Marktvolumen von Cloud Computing (B2B) in Deutschland nach Segment von 2011 bis 2015 und Prognose für 2020. Statista GmbH. Read on 04.05.2021. https://de.statista.com/statistik/daten/studie/168463/umfrage/prognose-zur-marktentwicklung-fuer-cloud-computing-in-deutschland/.

Steiner, R. 2021. Grundkurs Relationale Datenbanken. Einführung in die Praxis der Datenbankentwicklung für Ausbildung, Studium und IT-Beruf. 10th ed. Wiesbaden: Springer Vieweg.

Stephens, G. C. 1989. International Journal of Physical Distribution & Materials Management. 19th ed. Bingley: Emerald Publishing Limited.

van der Wardt, R. n.d. Scrum Board: Der Nutzen einer Aufgabentafel. Agile Scrum Group. Read on 10.05.2021. https://agilescrumgroup.de/scrum-board/.

Vogel-Heuser, B., Bauernhansl, T. & Hompel, M. ten 2020. Handbuch Industrie 4.0. Band 3: Logistik. 3rd ed. Berlin: Springer Vieweg.

Volkswagen AG 2019. Volkswagen and Amazon Web Services entwickeln Industrial Cloud. Read on 12.04.2021. https://www.volkswagenag.com/de/news/2019/03/volkswagen-and-amazon-web-services-to-develop-industrial-cloud.html.

[redacted]

Volkswagen AG 2020b. Volkswagen bringt weitere Partnerunternehmen in die Industrial Cloud. Read on 12.04.2021. https://www.volkswagenag.com/de/news/2020/07/Industrial_Cloud.html.

Volkswagen AG 2021a. Logistics Data Lake. Group Wiki entry. Unpublished.

Volkswagen AG 2021b. Volkswagen Commercial Vehicles Range Europe. VW CV Marketing Database. Unpublished.

Volkswagen AG 2021c. Volkswagen Nutzfahrzeuge. Read on 10.05.2021. https://www.volkswagenag.com/de/brands-and-models/volkswagen-commercial-vehicles.html.

Volkswagen AG n.d.a. Die Volkswagen PKI. Group Wiki entry. Unpublished.

[redacted].

[redacted].

Volkswagen AG n.d.d. Portrait & Produktionsstandorte. Read on 12.04.2021. https://www.volkswagenag.com/de/group/portrait-and-production-plants.html.

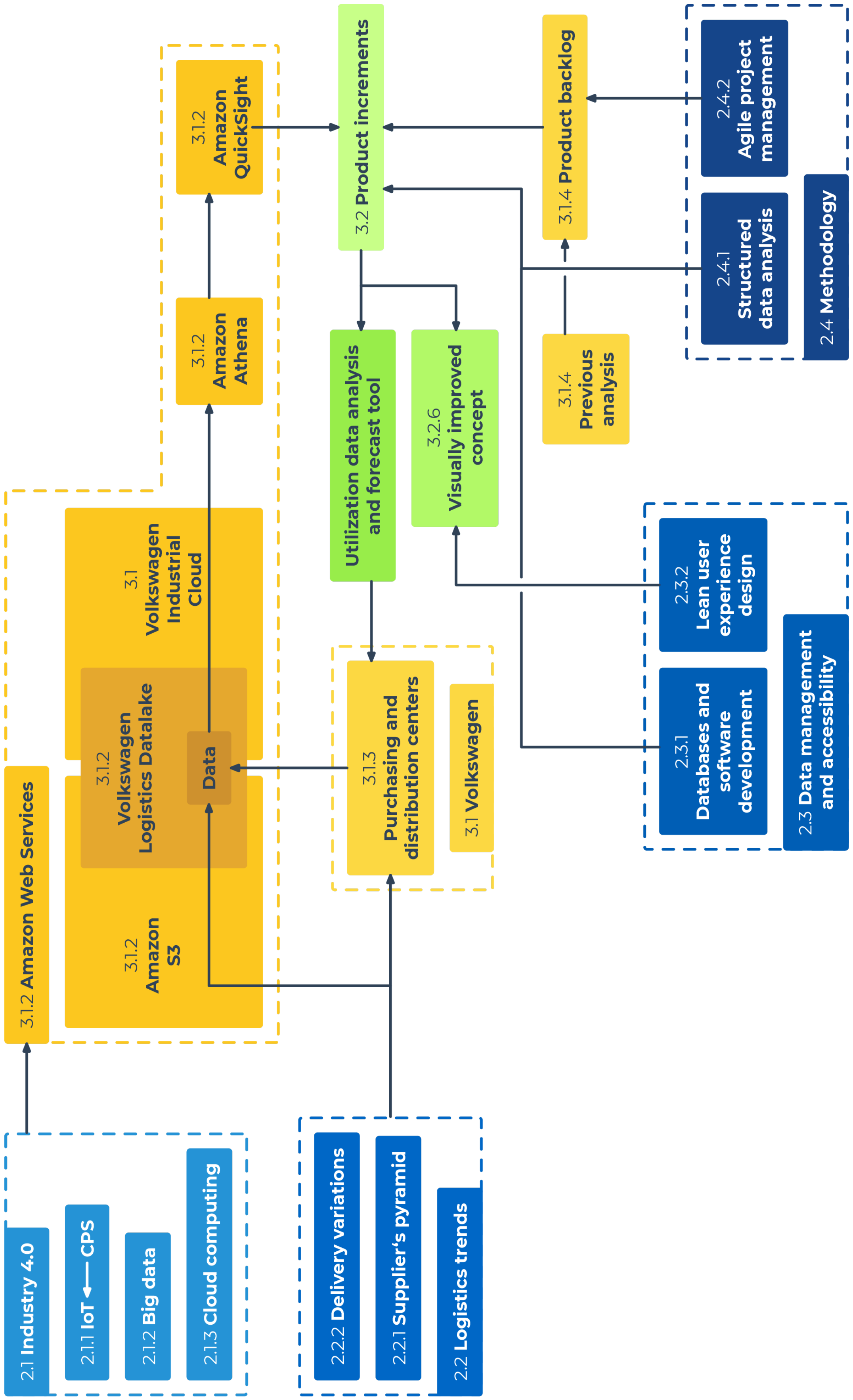
Weber, J. 2012. Logistikkostenrechnung. Kosten-, Leistungs- und Erlösinformationen zur erfolgsorientierten Steuerung der Logistik. 3rd ed. Heidelberg: Springer Vieweg.

Ximea n.d. Industry 4.0. Read on 12.04.2021. https://www.ximea.com/support/projects/vision-libraries/wiki/Industry_40.

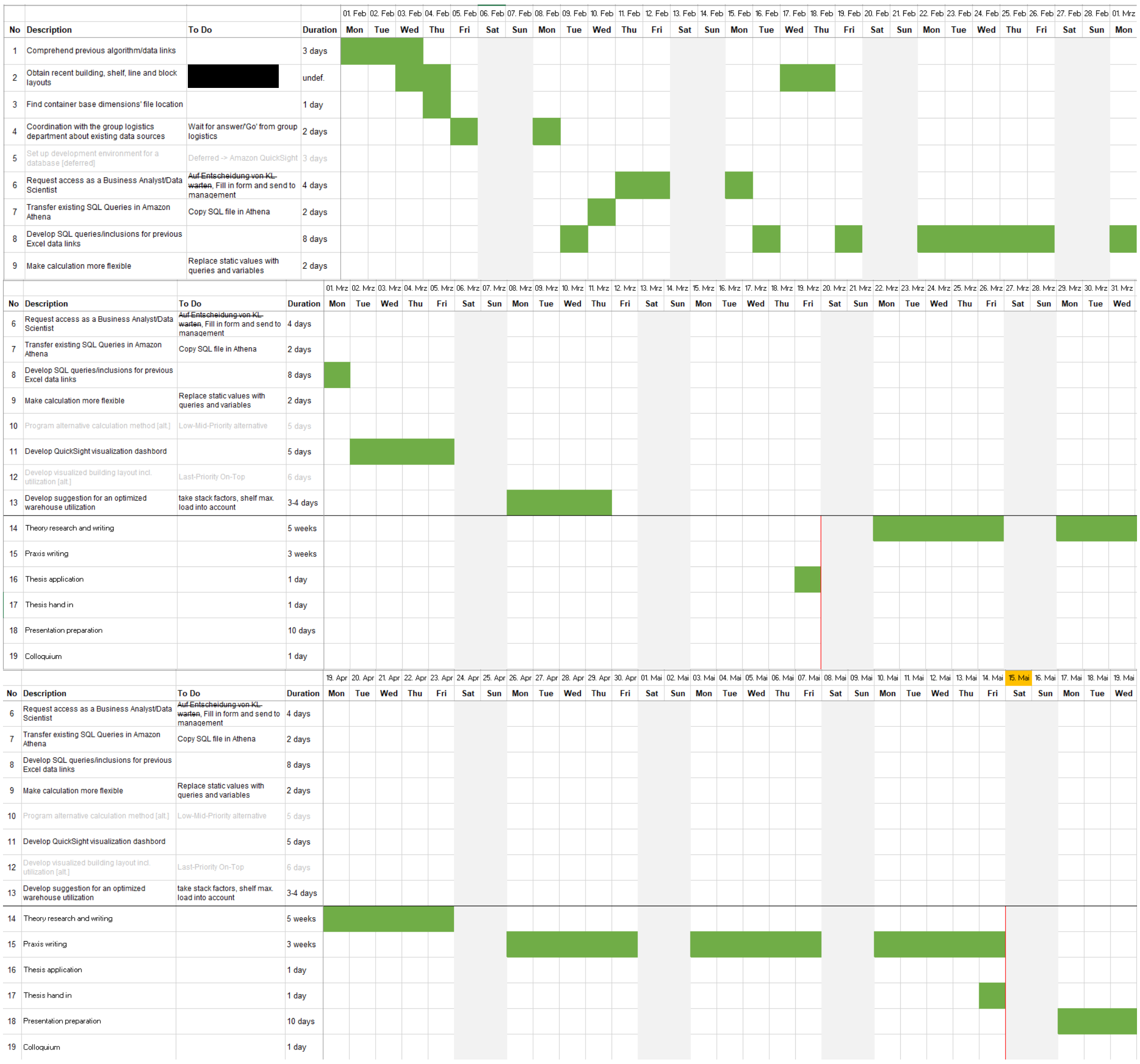
Yampolsky, A. 2018. The Principle of Least Astonishment. UX Planet. Read on 12.04.2021. https://uxplanet.org/the-principle-of-least-astonishment-bc3f67991510.

APPENDICES

Appendix 1. Thesis topic overview


2.1 Industry 4.0
2.1.1 IoT ⟵ CPS
2.1.2 Big data
2.1.3 Cloud computing
2.2 Logistics trends
2.2.1 Supplier's pyramid
2.2.2 Delivery variations
3.1.2 Amazon Web Services
3.1.2 Amazon S3
3.1.2 Volkswagen Logistics Datalake
Data
3.1 Volkswagen Industrial Cloud
3.1.2 Amazon Athena
3.1.2 Amazon QuickSight
3.1.3 Purchasing and distribution centers
3.1 Volkswagen
Utilization data analysis and forecast tool
3.2.6 Visually improved concept
3.2 Product increments
3.1.4 Product backlog
3.1.4 Previous analysis
2.3 Data management and accessibility
2.3.1 Databases and software development
2.3.2 Lean user experience design
2.4 Methodology
2.4.1 Structured data analysis
2.4.2 Agile project management

Appendix 2. Gantt project planning charts

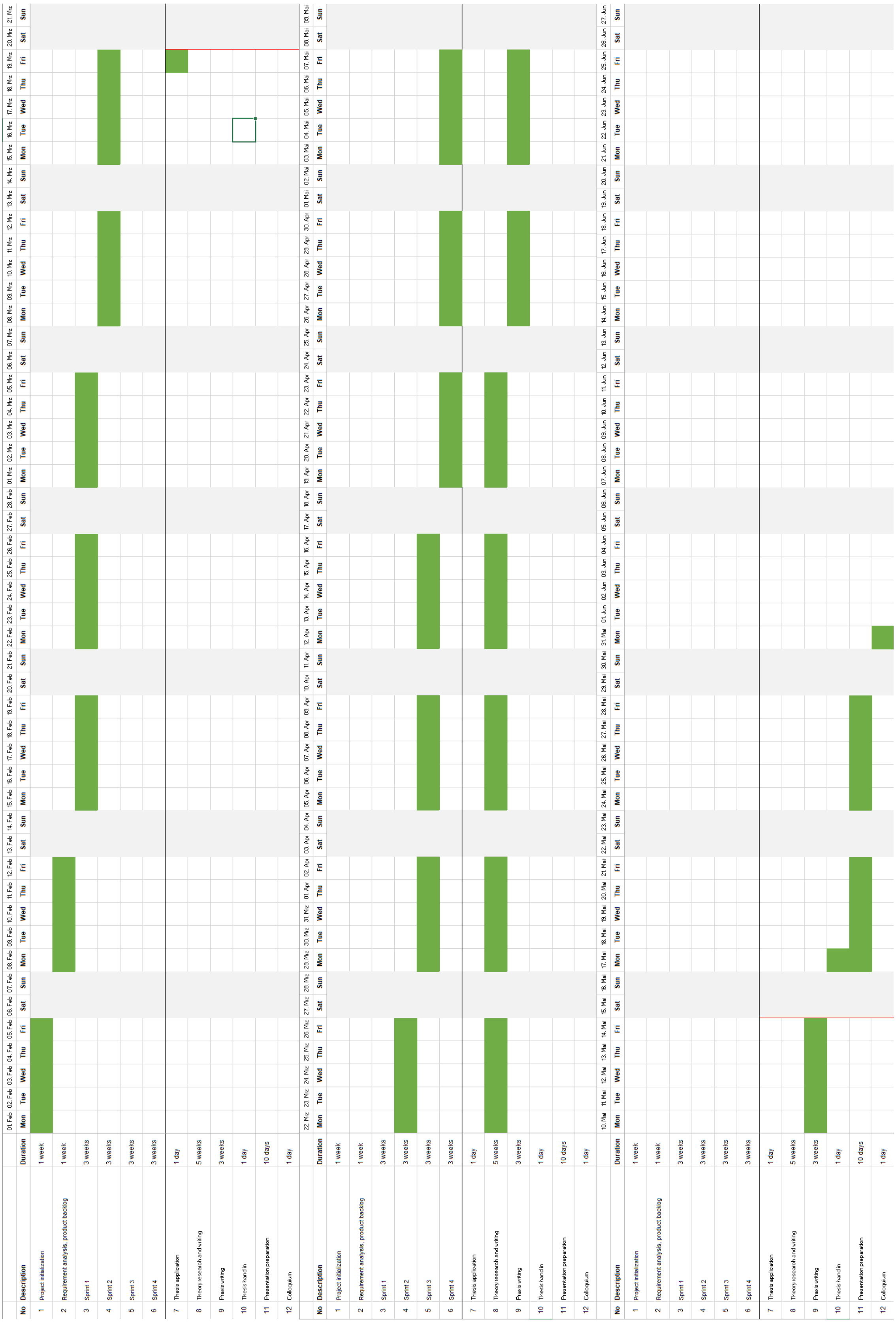


				01. Feb	02. Feb	03. Feb	04. Feb	05. Feb	06. Feb	07. Feb	08. Feb	09. Feb	10. Feb	11. Feb	12. Feb	13. Feb	14. Feb	15. Feb	16. Feb	17. Feb	18. Feb	19. Feb	20. Feb	21. Feb	22. Feb	23. Feb	24. Feb	25. Feb	26. Feb	27. Feb	28. Feb	01. Mrz
No	Description	To Do	Duration	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon
1	Comprehend previous algorithm/data links		3 days	🟩	🟩	🟩																										
2	Obtain recent building, shelf, line and block layouts	[redacted]	undef.			🟩	🟩													🟩	🟩											
3	Find container base dimensions' file location		1 day				🟩																									
4	Coordination with the group logistics department about existing data sources	Wait for answer/'Go' from group logistics	2 days					🟩			🟩																					
5	Set up development environment for a database [deferred]	Deferred -> Amazon QuickSight	3 days																													
6	Request access as a Business Analyst/Data Scientist	~~Auf Entscheidung von KL warten~~, Fill in form and send to management	4 days											🟩	🟩			🟩														
7	Transfer existing SQL Queries in Amazon Athena	Copy SQL file in Athena	2 days										🟩																			
8	Develop SQL queries/inclusions for previous Excel data links		8 days									🟩							🟩			🟩			🟩	🟩	🟩	🟩	🟩			🟩
9	Make calculation more flexible	Replace static values with queries and variables	2 days																													

				01. Mrz	02. Mrz	03. Mrz	04. Mrz	05. Mrz	06. Mrz	07. Mrz	08. Mrz	09. Mrz	10. Mrz	11. Mrz	12. Mrz	13. Mrz	14. Mrz	15. Mrz	16. Mrz	17. Mrz	18. Mrz	19. Mrz	20. Mrz	21. Mrz	22. Mrz	23. Mrz	24. Mrz	25. Mrz	26. Mrz	27. Mrz	28. Mrz	29. Mrz	30. Mrz	31. Mrz
No	Description	To Do	Duration	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed
6	Request access as a Business Analyst/Data Scientist	~~Auf Entscheidung von KL warten~~, Fill in form and send to management	4 days																															
7	Transfer existing SQL Queries in Amazon Athena	Copy SQL file in Athena	2 days																															
8	Develop SQL queries/inclusions for previous Excel data links		8 days	🟩																														
9	Make calculation more flexible	Replace static values with queries and variables	2 days																															
10	Program alternative calculation method [alt.]	Low-Mid-Priority alternative	5 days																															
11	Develop QuickSight visualization dashbord		5 days		🟩	🟩	🟩	🟩																										
12	Develop visualized building layout incl. utilization [alt.]	Last-Priority On-Top	6 days																															
13	Develop suggestion for an optimized warehouse utilization	take stack factors, shelf max. load into account	3-4 days								🟩	🟩	🟩	🟩																				
14	Theory research and writing		5 weeks																						🟩	🟩	🟩	🟩	🟩			🟩	🟩	🟩
15	Praxis writing		3 weeks																															
16	Thesis application		1 day																		🟩													
17	Thesis hand in		1 day																															
18	Presentation preparation		10 days																															
19	Colloquium		1 day																															

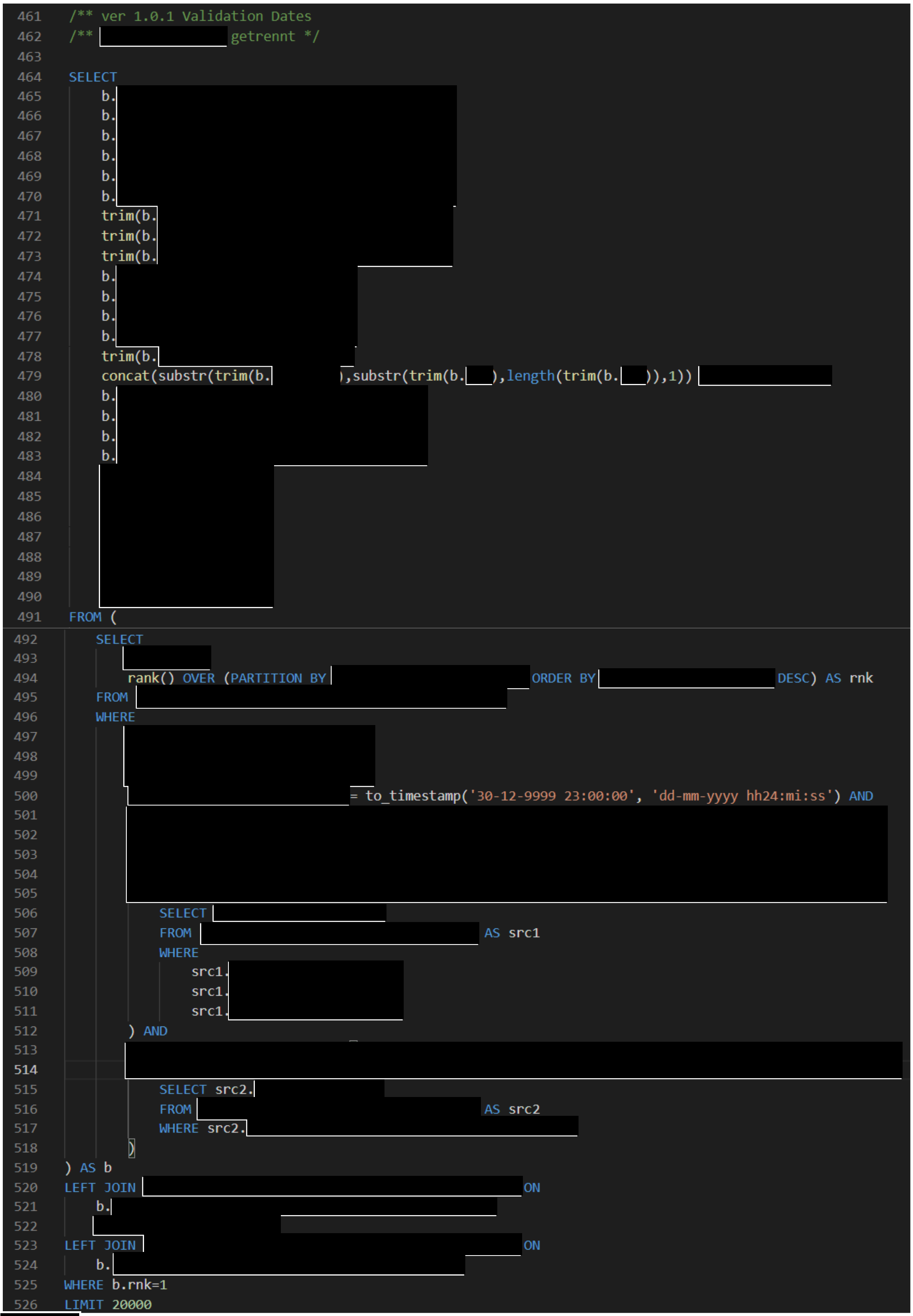
				19. Apr	20. Apr	21. Apr	22. Apr	23. Apr	24. Apr	25. Apr	26. Apr	27. Apr	28. Apr	29. Apr	30. Apr	01. Mai	02. Mai	03. Mai	04. Mai	05. Mai	06. Mai	07. Mai	08. Mai	09. Mai	10. Mai	11. Mai	12. Mai	13. Mai	14. Mai	15. Mai	16. Mai	17. Mai	18. Mai	19. Mai
No	Description	To Do	Duration	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed
6	Request access as a Business Analyst/Data Scientist	~~Auf Entscheidung von KL warten~~, Fill in form and send to management	4 days																															
7	Transfer existing SQL Queries in Amazon Athena	Copy SQL file in Athena	2 days																															
8	Develop SQL queries/inclusions for previous Excel data links		8 days																															
9	Make calculation more flexible	Replace static values with queries and variables	2 days																															
10	Program alternative calculation method [alt.]	Low-Mid-Priority alternative	5 days																															
11	Develop QuickSight visualization dashbord		5 days																															
12	Develop visualized building layout incl. utilization [alt.]	Last-Priority On-Top	6 days																															
13	Develop suggestion for an optimized warehouse utilization	take stack factors, shelf max. load into account	3-4 days																															
14	Theory research and writing		5 weeks	🟩	🟩	🟩	🟩	🟩																										
15	Praxis writing		3 weeks								🟩	🟩	🟩	🟩	🟩			🟩	🟩	🟩	🟩	🟩			🟩	🟩	🟩	🟩	🟩					
16	Thesis application		1 day																															
17	Thesis hand in		1 day																									🟩					
18	Presentation preparation		10 days																														🟩	🟩	🟩
19	Colloquium		1 day																															

Initial Gantt chart (1/2)



Updated Gantt chart (2/2)

Appendix 3. Final SQL queries



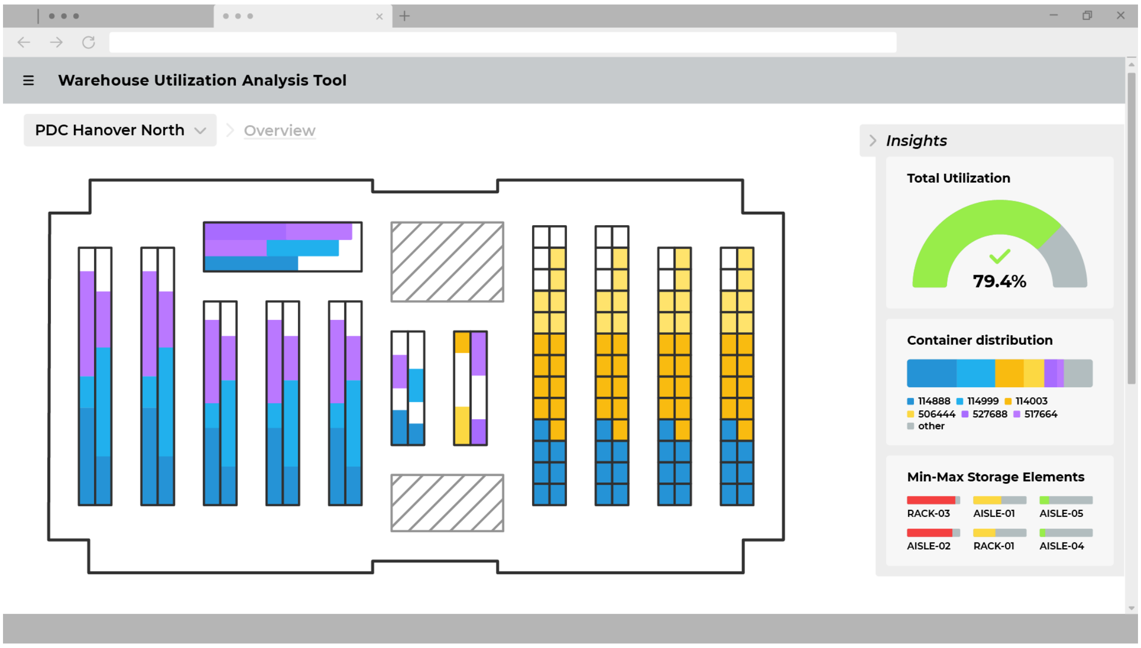
```
/** ver 1.0.1 Validation Dates
/** [redacted] getrennt */

SELECT
    b.[redacted]
    b.
    b.
    b.
    b.
    b.
    trim(b.
    trim(b.
    trim(b.
    b.
    b.
    b.
    b.
    trim(b.
    concat(substr(trim(b.[redacted]),substr(trim(b.[redacted]),length(trim(b.[redacted])),1)) [redacted]
    b.[redacted]
    b.
    b.
    b.







FROM (
    SELECT
        [redacted]
        rank() OVER (PARTITION BY [redacted] ORDER BY [redacted] DESC) AS rnk
    FROM [redacted]
    WHERE
        [redacted]


        = to_timestamp('30-12-9999 23:00:00', 'dd-mm-yyyy hh24:mi:ss') AND
        [redacted]




            SELECT [redacted]
            FROM [redacted] AS src1
            WHERE
                src1.[redacted]
                src1.
                src1.
        ) AND
        [redacted]

            SELECT src2.[redacted]
            FROM [redacted] AS src2
            WHERE src2.[redacted]
        )
) AS b
LEFT JOIN [redacted] ON
    b.[redacted]
    [redacted]
LEFT JOIN [redacted] ON
    b.[redacted]
WHERE b.rnk=1
LIMIT 20000
```

[redacted] query

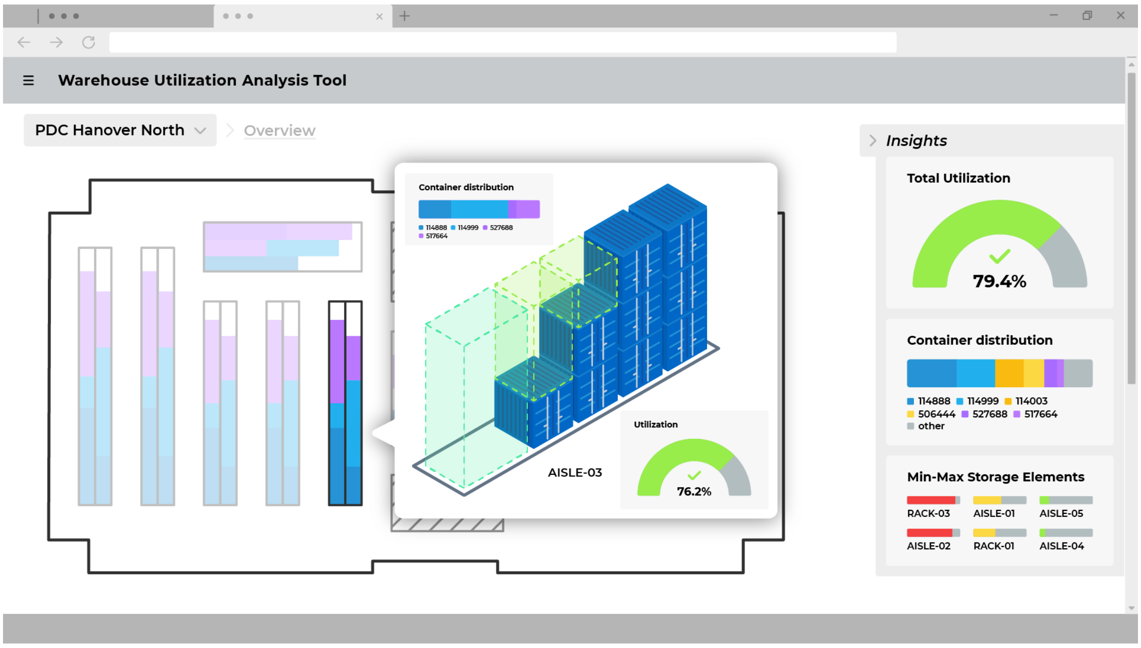
```
/* [redacted] */

SELECT [redacted]
FROM [redacted]
WHERE [redacted]
LIMIT 20000

/* [redacted] */

SELECT [redacted]
FROM [redacted]
WHERE [redacted]
LIMIT 20000
```

[redacted] queries

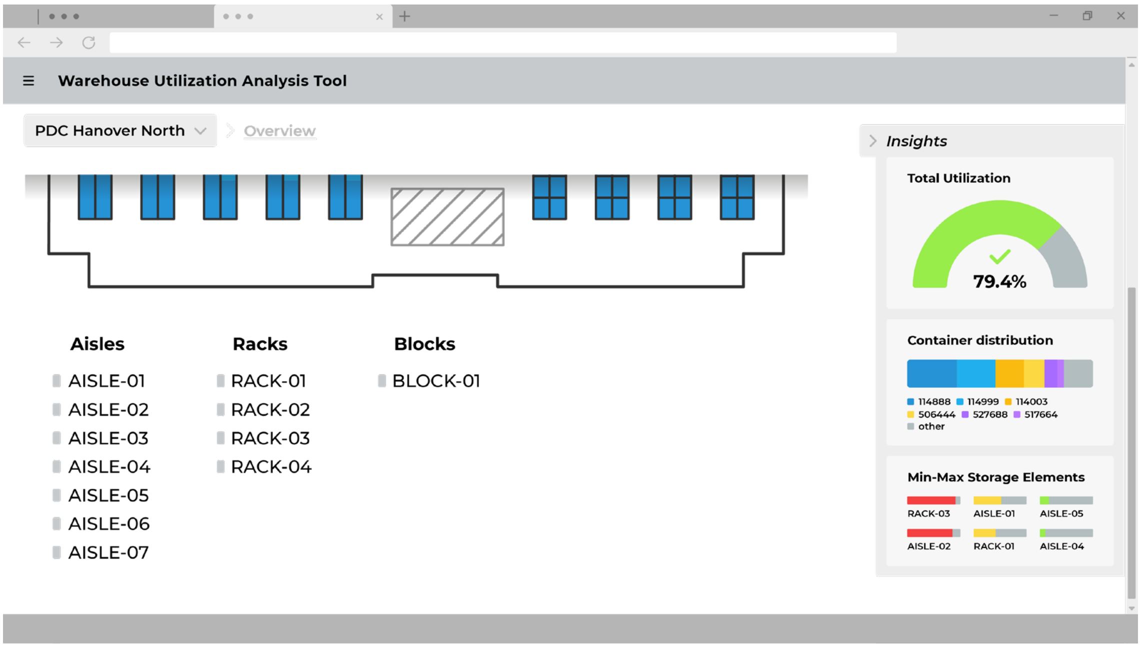
Appendix 4. Visualization concept



Overview (1/3)



Single storage element insight on hover event (2/3)



Warehouse Utilization Analysis Tool

PDC Hanover North > Overview

Aisles

- AISLE-01
- AISLE-02
- AISLE-03
- AISLE-04
- AISLE-05
- AISLE-06
- AISLE-07

Racks

- RACK-01
- RACK-02
- RACK-03
- RACK-04

Blocks

- BLOCK-01

Insights

Additional list of storage elements below (3/3)