DOCTORAL THESIS

Paint film quality predictive model development in an automotive paint shop

Author:
Javier Salcedo Hernández

Supervisors:
Dr. Jon García Barruetabeña
Dr. Iker Pastor López

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Engineering for the Information Society and Sustainable Development

in the

University of Deusto

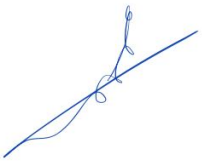
November 23, 2021

DOCTORAL THESIS

Paint film quality predictive model development in an automotive paint shop

Supervisor
Jon García Barruetabeña

Supervisor
Iker Pastor López



Author
Javier Salcedo Hernández

Dedicado a mis padres.

Por tanto. Por todo.

Declaration of Authorship

I, Javier Salcedo Hernández, declare that this thesis titled, "Paint film quality predictive model development in an automotive paint shop" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

"Past experience carries with its advantages the drawback that things never happen the same way again."

Winston S. Churchill

Abstract

This doctoral thesis studies the requirements and proposes the necessary procedures for the development of a predictive model of product quality in an automotive paint shop.

The requirements identified and the proposed procedures for the development of the predictive model have been divided into three parts, deployed sequentially. The first part, definition, is focused on the theoretical needs of the identification of the variables that will serve to train the predictive model. The second part, generation, is focused on the requirements and procedures that lead to the creation of a valid data set with the values of the variables defined in the previous step. In the third part, analysis, the requirements that said data must meet in order to train a model effectively are described. Then, the appropriate actions, based on advanced analytics, that are necessary for the development of the predictive solution are defined.

Once the requirements were identified and the theoretical procedures were developed for the development of the predictive model, an application case was carried out in a real automotive paint shop. In this experiment, the limits of the theoretical proposal have been identified and methodologies based on the implementation of the state of the art applied to this particular case have been proposed so that they lead to surpassing said limits.

Resumen

Este documento es una memoria de Tesis presentada en cumplimiento de los requerimientos para la obtención del grado de Doctor. La investigación titulada 'Paint film quality predictive model development in an automotive paint shop' estudia los requisitos y plantea los procedimientos necesarios para el desarrollo de un modelo predictivo de calidad de producto en una planta de pintura de automoción.

Los requisitos identificados y los procedimientos propuestos para el desarrollo del modelo predictivo se han dividido en tres partes, desplegadas de forma secuencial. La primera parte, definición, está enfocada en las necesidades teóricas de la identificación de las variables que van a servir para entrenar el modelo predictivo. La segunda parte, generación, está centrada en los requisitos y procedimientos que den lugar a la creación de un set de datos que contenga los valores de las variables definidas en el paso anterior. En la tercera parte, análisis, se describen los requisitos que deben cumplir los datos para poder entrenar un modelo predictivo de manera efectiva y se definen los pasos, basados en analítica avanzada, para el desarrollo de la solución predictiva.

Una vez se identificaron los requisitos y se desarrollaron los procedimientos teóricos para el desarrollo del modelo predictivo, se realizó un caso de aplicación en una planta de pintura de automoción real. En este experimento, se han identificado los límites de la propuesta teórica y se han propuesto metodologías basadas en la implementación del estado del arte aplicado a este caso particular que lleven a superar dichos límites.

Acknowledgements

First of all, I want to thank all those who in some way, no matter how indirectly that person thinks it may be, have contributed to the completion of this doctoral thesis.

More specifically, I want to thank, in no order of importance, the following:

To the Applied Mechanics group, especially Fernando.

To my directors Jon and Iker for enduring the unbearable, if you have survived this penance, the rest of your career is just enjoying yourself. Many thanks also to Iñaki and Borja for sharing part of your time to get all this started and running.

To all the doctoral students I met at Deustotech, especially Héctor, Adrián and Tony. Congratulations on your PhDs.

To my coffee mates Darlis, Ivan and Diego for entertaining me during my office hours, I hope you enjoyed those moments and conversations as much as I did.

To the Mercedes Benz staff of Vitoria and Sindelfingen, especially Gerardo and Javi, for believing in all this madness.

To Alberto and Olaia because you are already part of the story of my life. I hope our friendship will last for many many years.

To Soledad for accompanying me in part of this adventure.

To Ondiz for walking by my side along this path. Ondiz, I promise you that I will try not to be so stubborn. Also, thank you for your understanding when I was not willing to take a break with you because I was stuck modifying a sentence of this document for hours.

And to my parents, without whom nothing I have achieved in life would have been possible.

Contents

List of Figures

List of Tables

List of Abbreviations

ADS	**A**ggregierten **D**aten **S**treaming
BIW	**B**ody **I**n **W**hite
ERP	**E**nterprise **R**esource **P**lanning
ETL	**E**xtract **T**ransform **L**oad
IDE	**I**ntegrated **D**evelopment **E**nvironment
KPI	**K**ey **P**erformance **I**ndicator
KTL	**K**athodische **T**auch **L**ackierung
LFK	**L**ackierung **F**ehler **K**ontrolle
LFM	**L**ack **F**ailure **M**arking
LPD	**L**ackiert **P**rozess **D**iagnose
MDS	**M**ulti **D**imensional **S**caling
MPC	**M**odel **P**redictive **C**ontrol
PDS	**P**rozess **D**aten **S**treaming
PiA	**P**rozess **I**nformation **A**nalyse
PLC	**P**rogrammable **L**ogic **C**ontroller
SEC/KSO	**S**/**E**/**C** **K**lasse **K**onservieren **O**berfläche

CHAPTER 1

Introduction

THE new dynamics in the mobility market are driving a revolution in the automotive industry. To ensure competitiveness, vehicle manufacturers are expected to take a step forward in various engineering activities. Thus, the automotive industry must be more flexible, it must take advantage of the development of new technologies to improve production processes, generating less waste and pollution, increasing the perceived quality of the product and reducing production costs.

The digitization of production processes is the way to achieve these objectives. However, there is no clear way to apply this process. Especially in complex and highly specialized environments such as vehicle manufacturers. In the vehicle production factory, the painting process is especially critical both due to its complexity and its impact on the product quality perceived by the customer. Thus, the painting process is a bottleneck in many vehicle assembly plants. Despite its impact, the body painting process has remained unchanged for decades, depending on the experience of the workers and the limitations imposed by the process providers. For this reason, it is imperative to develop predictive control methodologies capable of ensuring the established quality standards.

1.1 The car body painting process

The automotive manufacturing process is divided into three main tasks. Firstly, in the body shop, the metallic structure of the vehicle is created. Secondly, in the paint shop, a corrosion prevention layer, a coloured layer and a bright protective layer are applied to the metallic surface. Finally, in the final assembly, the powertrain components, drivetrain elements and driver comfort equipment are installed to obtain the finished product. Painting is the most delicate step in vehicle manufacturing. A paint shop is a manufacturing bottleneck in many plants due to complexity of the car body painting process, tight production management labours and rigorous quality requirements. As defined above, the paint film is composed of a superposition of layers that are applied to the metallic surface throughout the painting process. This is done via a series of world wide standardized activities that are performed consecutively. Historically, these activities have been Streitberger and Dossel, 2008: washing the metal surface to remove dirt and metal remains from the body shop,

phosphating as metal surface pretreatment, cathodic electrodeposition of an anticorrosion layer, sealing and underbody protection, application of a preparation layer or primer surfacer, application of coloured enamel, application of a protective varnish layer and waxing. Along the process, quality controls are performed, normally to random car bodies, so the process is therefore controlled and adjustments are made due to, for example, each layer thickness variations or repetitive appearance of defects.

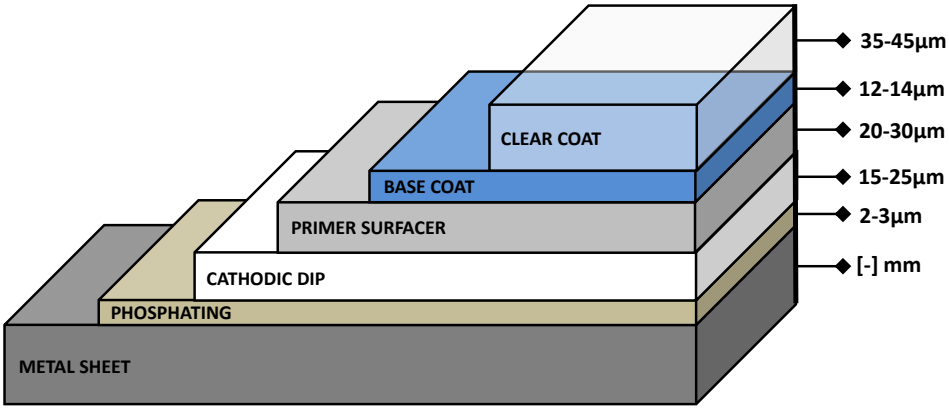


FIGURE 1.1: Paint layers that make up the paint film of the car body.

The thickness of the metal sheet depends on the area of the body and the function of the piece. The thickness used is usually from 0.4 mm to 0.8 mm in outer sheets of the vehicle body and from 0.8 mm to 1.5 mm in safety frames and reinforcements. The most common sizes are gauges 18, 20 and 22, which correspond to 1.2 mm, 0.9 mm and 0.7 mm thicknesses respectively (Materials, 2019).

An exhaustive control of the painted body is also carried out after applying the varnish so imperfections in the paint film layer can be corrected ensuring compliance with the paint shop quality criteria. This reworks are often required and can significantly increase the manufacturing costs. To help evolve the vehicle manufacturing process, the automotive industry has always been following the latest evolutions of technology as manufacturing software (Rambabua et al., 2018), manufacturing equipment ("Design optimisation of automotive component through numerical investigation for additive manufacturing"), materials (Cole and Sherman, 1995–Kuziak, Kawalla, and Waengler, 2008), tools (Garavaglia et al., 2019) or automated systems (Wollschlaeger, Sauter, and Jasperneite, 2017). In the paint shop, the development of coating machines, robots and automation makes craftsmanship no longer needed. This early processes digitisation makes automotive industry be inclined to adopt the proposals made by new technological paradigms as European Industry 4.0 Hermann, Pentek, and Otto, 2015 in order to increase process efficiency while maintaining product quality.

Quality, in a paint shop product, refers to corrosion protection and paint film layer final appearance and long lasting ability. Related to the final appearance, many different quality issues can be detected (*Paint Defects*). Among these, inclusions in the paint film (lifts in the paint film caused by the presence of a underlying contaminating element such as dirt or dust) and imperfections in the shape of craters (usually, as round depressions in the paint film) are the most common. These imperfections are due to the combination of multiple reasons, changes in air flows that allow to

approach impurities contained in the environment to the body, changes in temperatures and humidity that change the speed of evaporation of solvents, wear of paint application elements and others (*Paint Defects*). Although the process is profusely controlled and known, analysing the origin of each defect is a difficult task to perform because, on the one hand, they are usually the result of combinations of different factors even though process parameters are within the operating margins and, on the other hand, the defect cause can be only identified using destructive methods that can not be applied to every sample.

Historically, to detect and correct these defects, inspection zones have been implemented at the end of each relevant paint layer application. Here, specialized personnel carefully examine the surface of the body. The weak point of this method is the consistency of detection since human vision differs between individuals and, in addition, fatigues. Due to the advancement of artificial vision technology, these defect detection activities have been automated (Tornero et al., 2012) in such a way that they allow for constant evaluation criteria.

In this research, the quality criteria has been defined by using the data provided by an automatic fault detection process that has been applied to the entire production and in which the sensitivity has been constant throughout the study. This allows (applying the precepts proposed by Industry 4.0 (Bauernhansl, 2017) about analysis of large amounts of data) to venture into the development of solutions, such as predictive models, that allow improving the efficiency of processes. The aim of this research is the development of a predictive model that will enable to evolve from a corrective quality control to a preventive one, allowing to identify the factors that are going to lead to the appearance of a defect and being able to prevent it. Hence, the objective of this particular study is to verify the predictive capabilities of the paint process data on the appearance of defects in the paint film layer. The predictive models applied to the manufacturing industry are being applied mainly to predictive maintenance of equipment Peng, Dong, and Zuo, 2010 and logistics estimates and scheduling Mönch, Zimmermann, and Otto, 2006, movement of materials and products to optimize lots. The study to make the model, has been divided into three main steps. First, the definition, in which the quality criteria to be predicted are fixed and the relevant variables for the prediction are identified . Second, the generation of the data set, in which the data of the variables identified in the previous step are extracted and synchronized. Last but not least, the analytical part in which the model is developed. As a result, it is proved that the model shows predictive intelligence. Therefore, it can be concluded that coating process variables can predict the appearance of defects in the paint film layer.

1.2 Industrial-scientific problem

The painting process is made up of a number of complex operations that are carried out consecutively. Each process carried out may have influence on subsequent operations. Throughout the painting process, a limited number of vehicles are subjected to quality tests to ensure the correct painting of the paint film. However, these tests are corrective and do not guarantee the necessary target quality. If these quality tests detect a paint film defect, there is additional work on the chassis, which accrues an extra cost and can even lead to the disposal of the vehicle. As the chassis is not checked as it moves through the process, the paint problems accumulates and can create an even bigger problem. It is also not easy to guarantee a homogeneous

paint layer quality. But, couldn't the process be analysed so that we can guarantee the correct painting of the vehicles without having to check each and every one of them?. If the process can be understood, it can be controlled through the analysis of precess variables. And by controlling the process, it can be guaranteed with a high percentage of probability that the vehicles will be painted correctly and with the required quality. Understanding the process not only improves the final quality but also allows other aspects of the process to be analysed, such as energy consumption.

1.3 Outline of the thesis

The thesis is structured as follows:

Chapter 1: Introduction. The scientific and industrial problem is presented.

Chapter 2: State of the art. The literature is reviewed in search of methods to develop applications with data in a general way, then focusing on predictive models and specifically on those applied in industry. After this general review, cases of automotive paint plants in which an attempt has been made to apply a predictive model are studied. The chapter ends with a critical study of the state of the art and the hypothesis and objectives raised from it.

Chapter 3: Data set definition. A methodology is proposed to identify the variables that can describe the quality to be predicted and the process variables associated with them.

Chapter 4: Data set generation. A methodology is proposed for the generation of the data set that will be used for the development of a quality predictive model.

Chapter 5: Data set analysis. Data quality is checked against quality measurement criteria and a predictive model for paint quality is developed.

Chapter 6: Case study: Mercedes Benz Fábrica Vitoria. Procedures described in Chapters 3, 4 and 5 are applied to a real use case.

Chapter 7: Conclusions. The main conclusion of the thesis are summarised, the publications are presented and recommendations for future work are proposed.

CHAPTER 2

State of the Art

THE automotive manufacturing process is not a common topic when looking for research articles that are publicly available. This type of research related to the know-how of each automotive brand is their competitive advantage so they are reluctant to share it. In this case, it has been possible to analyse the paint shop of the Mercedes Benz plant in Sindelfingen (Germany), which is considered the most advanced in terms of paint shop process control and data extraction in the Daimler Group and therefore in the automotive world. Thus, its methods of process data extraction, data set generation and application development have been characterized.

The chapter starts with a description of general methodologies for the development of data related projects, so as they can be applied to the creation of a predictive model. Then, the most common steps to create this predictive model are defined: process data definition and data set generation methods will be described and then, finally, typical methods for modelling found in literature will be revised.

2.1 Methods for the development of data related projects

With digitization, it has been possible to extract a large and varied amount of data from the process. This makes the traditional way of handling information no longer valid. A new way of handling data is needed that allows to ask the right questions and get trustworthy answers. Old systems had to be completely designed before the arrival of the first data, they did not allow flexibility and they were not scalable with a greater arrival of information. It was also not easy to identify which data was valid to answer which question or, of course, to answer questions in real time, since the data had to be stored and retrieved from the system before being used. This has led to research on how to manage information. Some of the fields of research are:

- Data ingestion: develop a procedure to ensure how to obtain a large volume of data, regardless of the structure of each one.
- Integration, quality and enrichment of the data: how to know that the data read is valid for later use, integrate data of different types, etc.
- Discovery and consumption of data: how to recognize the data that we need and how to serve it.

- Data governance: how to guarantee the usability, security, integrity and availability of the data.

Some reference research groups in this field are: Data Research Center (University of Essex), IBM Big Data and Analytics Hub, Gartner and DAMA-International. Virtually every organization has its own solution for managing data in the company. Some of the most widespread environments (frameworks) are:

- MIKE 2.0. Method for an Integrated Knowledge Environment: it is a work system to develop best information management practices, linked to common business problems and specific technology solutions. Its scope spans the entire information supply chain within an organization, from how it is created, accessed, presented, and used in decision-making to how it is kept secure, stored, and destroyed.
- DAMA DMBoK v3. Data Management Body of Knowledge: it defines an industry-standard view of data management functions, terminology, and best practices, without detailing specific methods and techniques.
- COBIT 5. Control Objectives for Information and Related Technology: it is a best practices guide presented as a framework, aimed at the control and supervision of information technology (IT). It has a series of resources that can serve as a reference model for IT management, including an executive summary, a framework, control objectives, audit maps, tools for its implementation and mainly, a guide to management techniques.
- TOGAF. The Open Group Architecture Framework: it is an Enterprise Architecture scheme that provides an approach to the design, planning, implementation and governance of an enterprise information architecture. This architecture is generally modelled on four levels or dimensions: Business, Technology (IT), Data, and Applications. It has a set of base architectures that seek to facilitate the team of data architects how to define the current and future state of the architecture.
- ITIL v3. IT Infrastructure Library: It is made up of a series of books, each dedicated to a specific practice within IT management. The ITIL set of best practices provides a complete set of practices that covers not only technical and operational processes and requirements, but also relates to strategic management, operations management and financial management of a modern organization.
- DGI. Data Governance Framework: is a logical structure for classifying, organizing, and communicating the complex activities involved in making decisions about company data and taking action.
- EIM Framework. Enterprise Information Management: it is a set of business processes, disciplines and practices used to manage the information created from the data of an organization as a business asset. This framework ensures that high quality information is available, protected, and controlled.

2.1.1 Process data definition

No literature has been found on this topic. The examples that can be found in scientific documents or books, focus mainly on the application of statistical techniques and show a set of data already created, without showing the steps taken to form that set of data. Neither about how the data must be named, only automatically by means

of an algorithm, but it is not valid for the realization of this thesis. Neither data has been found about how to order the data or what characteristics the magnitudes that compose it must have.

That is why this part will be developed through experimentation day by day and according to the needs of the research.

2.1.2 Process data set generation

Despite the fact that the frameworks presented above propose data management methods to generate analytical solutions, none use industrial cases that reveal the difficulties and characteristics of generating an industrial data set. For this reason, a research stay was carried out at the Mercedes Benz plant in Sindelfingen (Germany) since this plant is the cutting edge in data generation in the paint shop process. The following objectives were set:

- Regarding the extraction of process data, the objectives were:
  - Analyze the data generation structure of the painting process at the Sindelfingen paint shop: that is, in what fundamental components the generation of data is divided; the processes and equipment that are the source of data in the paint plant.
  - Analyze the data generation architecture of the painting process at the Sindelfingen paint shop: here, it is intended to know the structure of systems that support the generation of data; the networks, processing or storage servers that are used to carry out the information management.
  - Analyze the data generation methods of the painting process at the Sindelfingen paint shop: this point, as the most relevant point of the process data extraction part, aims to analyse the methods used for the acquisition and transformation of data from the different sources identified in the plant.
- Regarding the synchronization of process data, the objectives were:
  - Analyze the methods of temporal synchronization of variables: learn how the Timestamps of the paint shop process variables are synchronized, when variables are relevant within the painting process.
  - Analyze the synchronization methods in terms of location of the variables: how the process variables are synchronized around a primary key, in this case, the location of each vehicle chassis as it circulates through the painting process.

Data Generation Structure in the Sindelfingen paint shop

As a result of the first of the objectives to be met in the research stay in Sindelfingen, the following procedure is defined, using Figure 2.1 as a reference, to specify the fundamental components of the painting process; in which processes, for which elements, through which point and with which method data are generated in the Sindelfingen paint plant. This allows knowing the singularities of the Sindelfingen process, classifying its elements in levels and showing their relationships.

Each level, starting with Level 1 - Process, which is the only one that only defines one option for each case, has one or more options associated with the next level. The

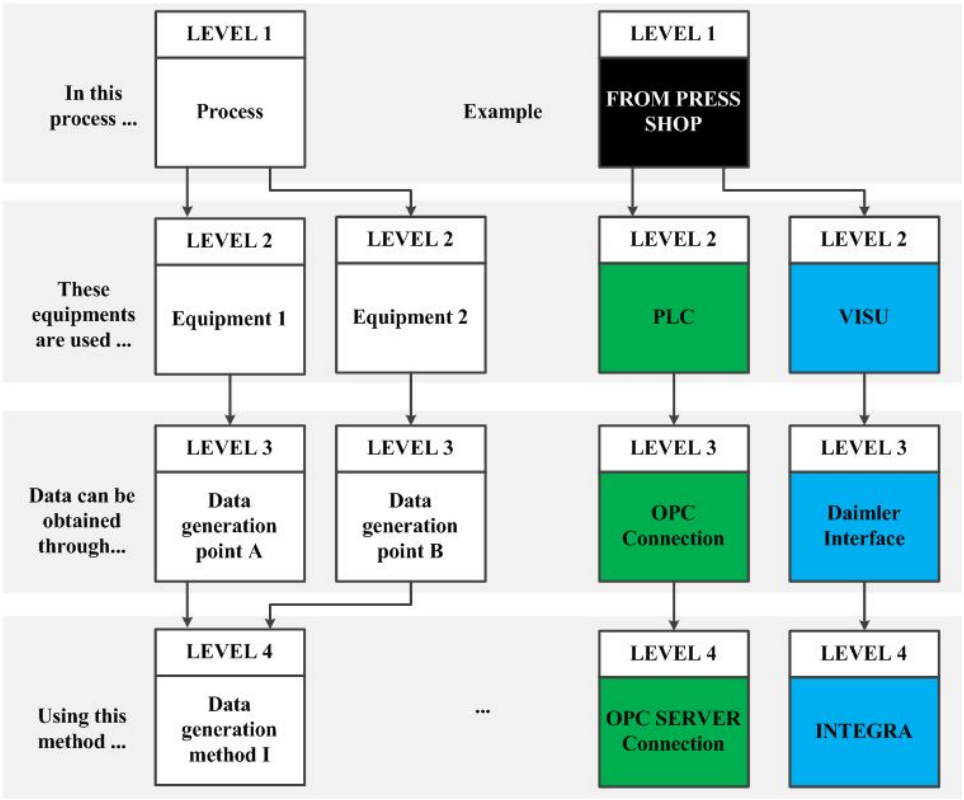


FIGURE 2.1: Procedure to follow the generation structure characterization.

general structure of data generation in Sindelfingen, combining all the variants of all the levels, is presented in Figure 2.2.

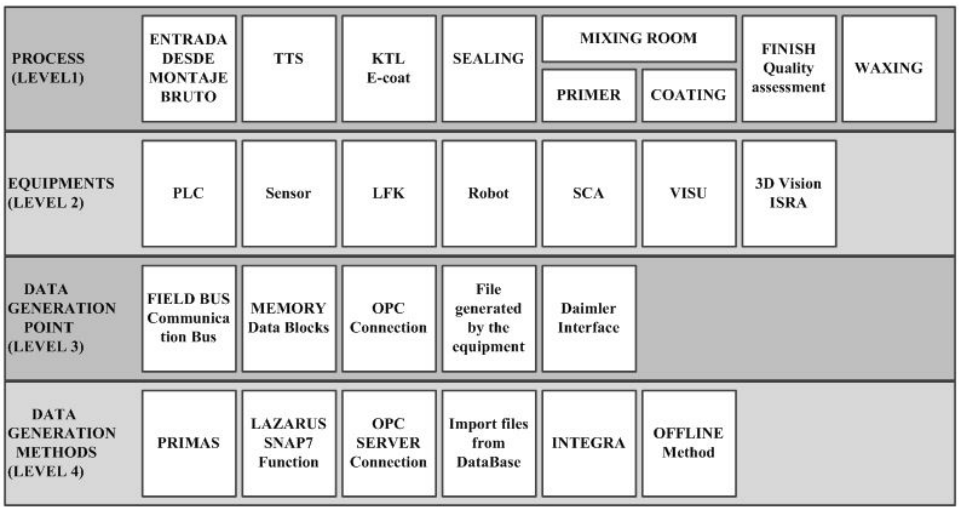


FIGURE 2.2: General structure of data generation in Sindelfingen.

In Figure 2.2 the nine options of Level 1 - Process, the seven options of Level 2 - Equipment, the five options of Level 3 - Generation point, and the six ooptions of Level 4 - Generation methods, are presented simultaneously. The levels presented are defined as:

- Level 1, process. Nine relevant parts are distinguished (see Figure 3) that are data sources in the painting process. The mixing room is considered as an independent process but is associated with the priming and finishing processes.
- Level 2, equipments. They are the devices or computers that have been identified as the main sources of data. Seven different ones are distinguished (see Figure 3) and each of them may contain variations in terms of suppliers or models. For instance:
  - PLC, which includes Siemens S7 PLC and Beckhoff PLC.
  - Sensor, which refers to any independent sensor not associated with any of the other equipment.
  - Robot, which includes KUKA robots and those related to the supplier DÜRR.

  These differentiations are not relevant when defining the generation structure since, for example, PLC from different suppliers share the same generation points and methods, etc. The same computer can have several data generation points depending on the type of data to be extracted. For example, PLC have the fieldbus as the generation point for real-time data and memory data blocks for control signals or alarms.
- Level 3, data generation point. They are the specific places of the equipment, or elements related to them, from which the data is acquired. Five possible generation points are defined (see Figure 3) that go from an internal element of the equipment such as internal memory blocks, to a communication bus of the equipment with an actuator or peripheral or to a file generated by the equipment. Different computers in the same process can have the same data generation point.
- Level 4, data generation method. It is the technique, extraction system or action carried out to obtain data from each point of generation. Six different methods are applied in the Sindelfingen plant and will be presented next.

Next are presented, for each Level 1 - Relevant process in which the body painting has been divided, the equipment relationships, generation points and methods used in each one of them. The particular case of the OFFLINE Method is not related to any of the processes since it is a solution used prior to the implementation of the rest of the methods.

- LEVEL 1 - ENTRANCE FROM PRESS SHOP

  As shown in Figure 2.3, the following data generation methods are used in this process:
  - In green, the connections to the PLC are shown using the OPC SERVER connection method (there are two connections of this type in this process).
  - In blue, the INTEGRA connection to generate data from the VISU (there is a connection in the process). The OPC method allows transferring data from non-input systems from raw assembly, such as the manufacturing plan, to the process control systems. Thus, manufacturing data is assigned to each new body. The INTEGRA method allows obtaining data from the display screen present in the process.

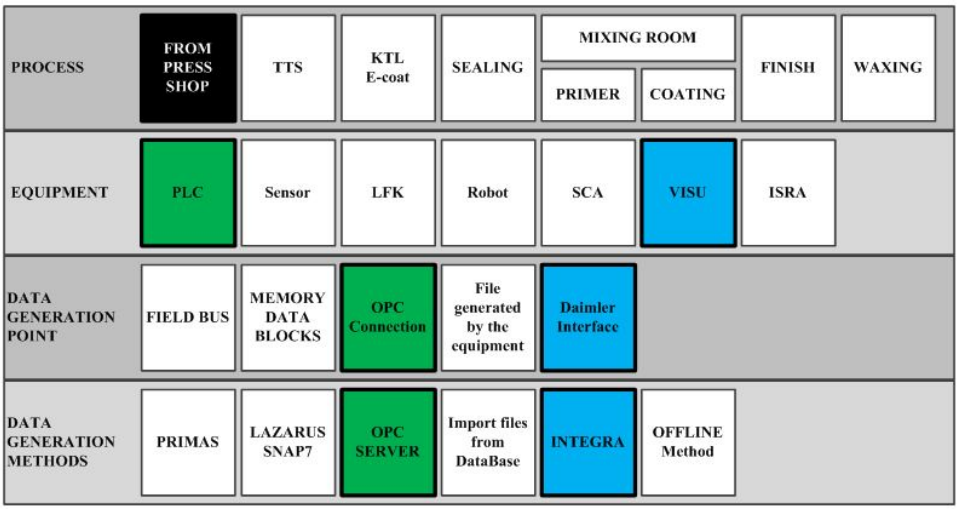


FIGURE 2.3: Generation structure at the entrance from the press shop.

- LEVEL 1 - TTS

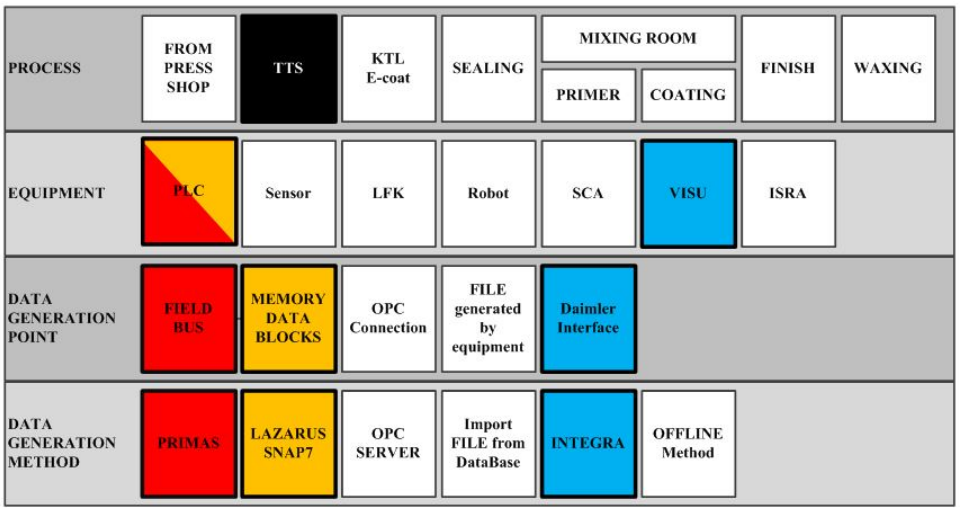


FIGURE 2.4: Generation structure in the TTS process.

As shown in Figure 2.4, the following data generation methods are used in this process:

  - In yellow, connections to PLC through LAZARUS SNAP7 (there are two connections).
  - In red, connections to PLC through PRIMAS (there are two connections).
  - In blue, INTEGRA connections for VISU (there are two connections).

PLC data is generated by two methods, PRIMAS for real-time data and LAZARUS SNAP7 for data from PLC memory blocks. INTEGRA is used to generate data from the VISU.

- LEVEL 1 - KTL

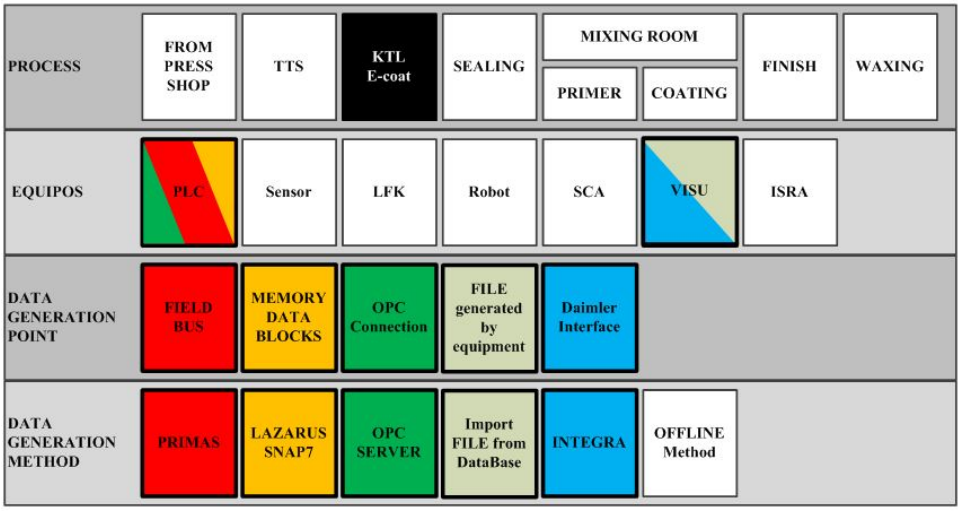


FIGURE 2.5: Generation structure in the KTL process.

As shown in Figure 2.5, it has the following connections:

- In yellow, connections to PLC through LAZARUS SNAP7 (2 connections).
- In red, connections to PLC through PRIMAS (2 connections).
- In green, connections to PLC through OPC SERVER (2 connections).
- In blue, INTEGRA connections for VISU (2 connections).
- In light green, Import of a file, csv file, from VISU.

In this case, PLC data is generated by three methods, PRIMAS is responsible for generating the data that circulates through the field bus that connects the controller with the devices. LAZARUS SNAP7 is in charge of reading data and signals from the PLC memory and OPC is in charge of writing data in the memory, such as an alarm calculated in a system outside the process.

- LEVEL 1 - SEALING

  As shown in Figure 2.6, this process has the following data generation methods:

  - In yellow, connections to PLC through LAZARUS SNAP7 (22 connections).
  - In blue, INTEGRA connections for VISU (22 connections).
  - In red, PLC connections through PRIMAS (22 connections).
  - In purple, connections to Robot through PRIMAS (180 connections).
  - In brown, PRIMAS connections for SCA (180 connections).
  - In turquoise green, PRIMAS connections for ISRA (90 connections).
  - In light green, import of log files generated by the ISRA system (90 file imports).

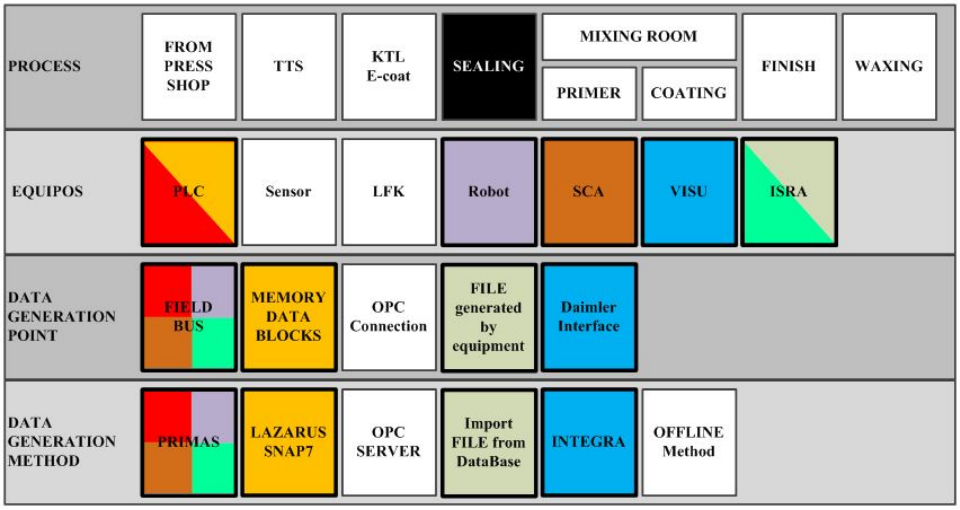


FIGURE 2.6: Generation structure in the sealing process.

The PRIMAS method is used to generate PLC, Robot, SCA and ISRA data. Robot and SCA form the SAM team for mastic application, but in this analysis both elements have been kept separately. Data present in the memory of the PLCs are also generated by LAZARUS SNAP7, VISU data by the INTEGRA method and files generated by the ISRA system are imported.

- LEVEL 1 - MIXING ROOM

  The fifth process involves generating data in the mixing room. It is separated from the primer and coating processes to treat the particular case of the connection of sensors directly to the PRIMAS system.

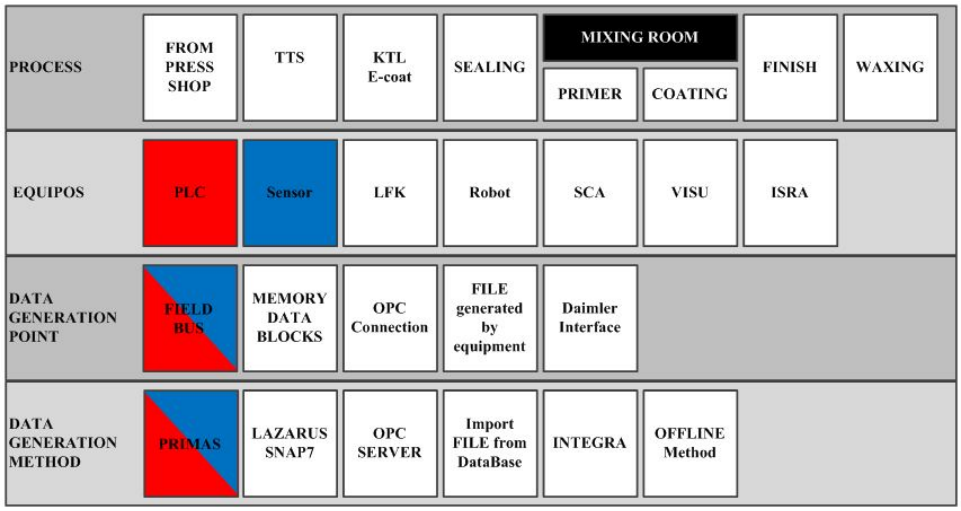


FIGURE 2.7: Generation structure in the mixing room.

As shown in Figure 2.7, it has the following connections:

  - In red, PRIMAS connections for PLC.

– In dark blue, data capture from independent sensors through their connection to the PRIMAS system.

In this process, it was necessary to have data from some independent pressure sensors to be able to extrapolate the viscosity of the paint material. It was decided to connect these sensors to the PRIMAS system's own bus, Interbus. Other process data is also obtained by connecting a PLC to the PRIMAS system.

- LEVEL 1 - PRIMER

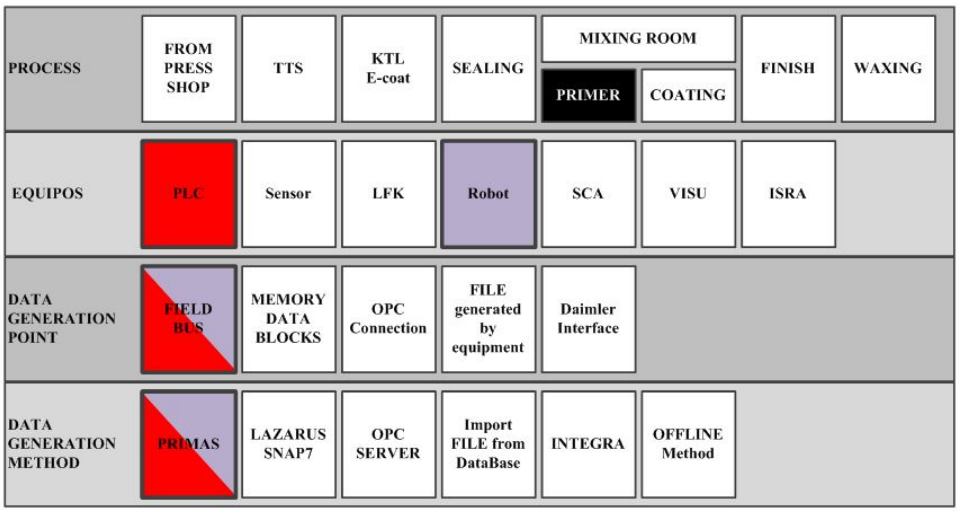


FIGURE 2.8: Generation structure in the primer application process.

As can be seen in Figure 2.8, this area has the following connections:

– In red, connection to PLC through PRIMAS (3 connections).

– In purple, connection to Robot through PRIMAS (2 connections).

Here, the main data generators are the PLC, distributed in the process in the functions of application control, cabin air conditioning and oven control. Data is also obtained from the robots present in the process.

- LEVEL 1 - COATING

As shown in Figure 2.9, the following data generation methods are available:

– In red, PRIMAS connections for PLC (24 connections).

– In purple, PRIMAS connections for Robot (18 connections).

– In green, OPC connection for PLC.

In this process, data are generated from the application control PLCs, the body transport system, the cabin air conditioning control and the drying oven (see Annex) through a connection to PRIMAS. Data from the DÜRR Application Robots is generated by PRIMAS and signals are communicated to the PLCs via the OPC connection.

- LEVEL 1 - FINISH

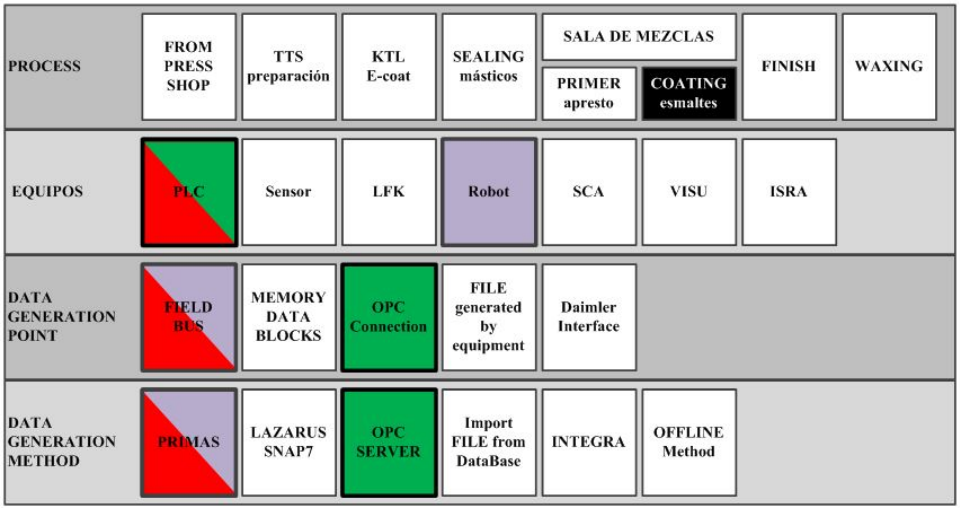


FIGURE 2.9: Generation structure in the coating application process.

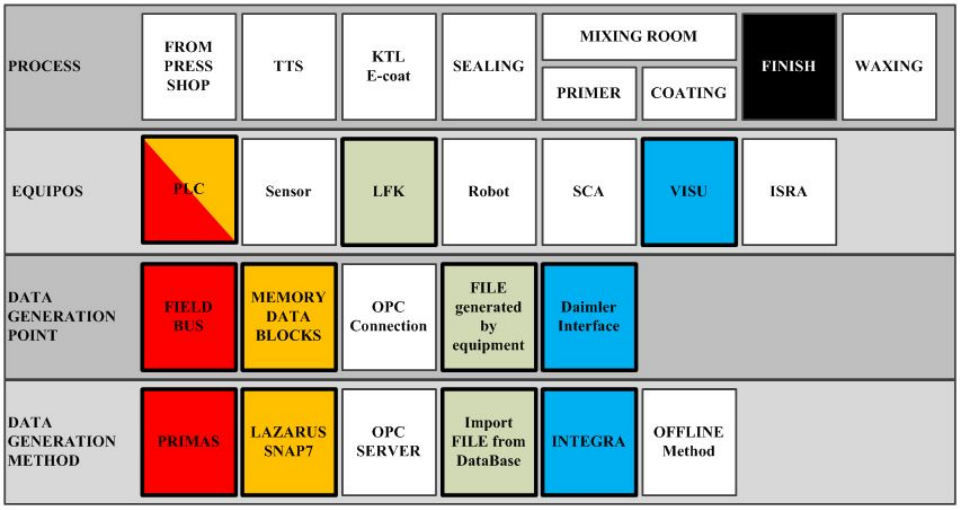


FIGURE 2.10: Generation structure in the quality assessment process, finish.

As can be seen in Figure 2.10, quality control has the following data generation methods:

- In red, connections to PLC through PRIMAS (3 connections).
- In blue, INTEGRA connections for VISU (3 connections).
- In light green, Import of an xml file with the defects detected by the LFK system.
- In yellow, connections to data blocks in PLC memory through LAZARUS SNAP7 (2 connections).

PLC data is generated using PRIMAS, VISU data with the INTEGRA method, data from the defect repair line that has two LAZARUS SNAP7 connections to the body transporter control PLC and air conditioning. The xml files generated by the LFK system are imported, with the detected defects.

- LEVEL 1 - WAXING

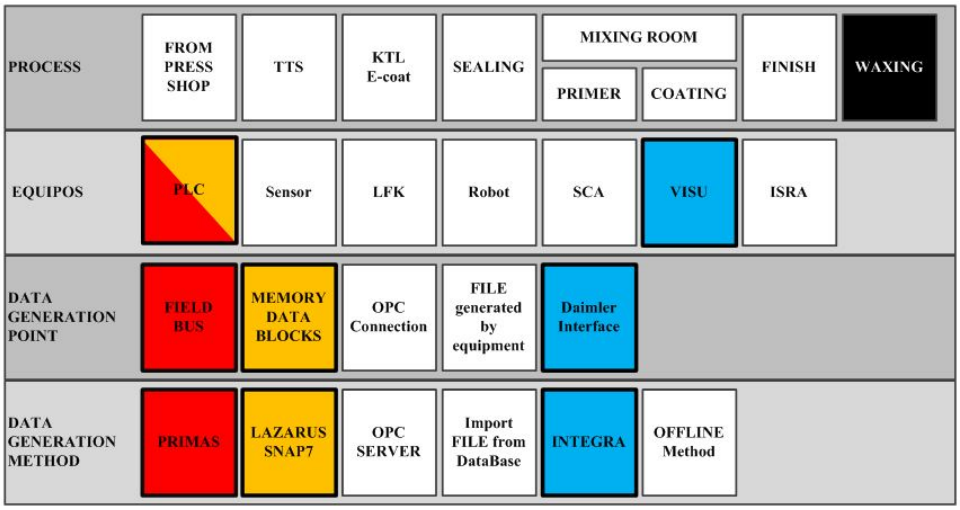


FIGURE 2.11: Generation structure in the waxing process.

As shown in Figure 2.11, the wax process has the following data generation methods: As can be seen in Figure 2.10, quality control has the following data generation methods:

- In red, connections to PLC through PRIMAS (4 connections).
- In yellow, connections to PLC through LAZARUS SNAP7 (4 connections).
- In blue, connections to VISU through INTEGRA (4 connections).

Data generation architecture in the Sindelfingen paint shop

As a continuation of the first objective of knowing the singularities of the Sindelfingen process and as an introduction to the data generation methods, this section presents the architecture of systems that support the generation of data in the Sindelfingen paint plant. That is, the communication networks, servers and other elements that allow to take the data from the generation methods to the supervision and visualization tools that generate value from the data. Figure 2.12 shows the most relevant elements in the data generation architecture at the Sindelfingen paint plant.

Two networks are distinguished: the data source equipment represented by the PLCs (each with its corresponding data generation point), the most important data generation methods, and the information visualization systems (that of the PRIMAS method and the plant data visualization system, PiA, defined in the annex to this document). PiA has services that are responsible for collecting data and making it available to plant users through a web application (Apache Web Server). Regarding networks, the top one (1Gbit) belongs to the IT department and is the network to which the plant control equipment (PLC) is connected.

The second network (10Gbit), called the LOCAL network and located at the bottom, belongs to the department responsible for the process data. This second network is motivated by the bandwidth needs required to manage the generated process data (during the development of the architecture, the IT network was unable to support all the data traffic). In reference to the data source equipment, each PLC has its

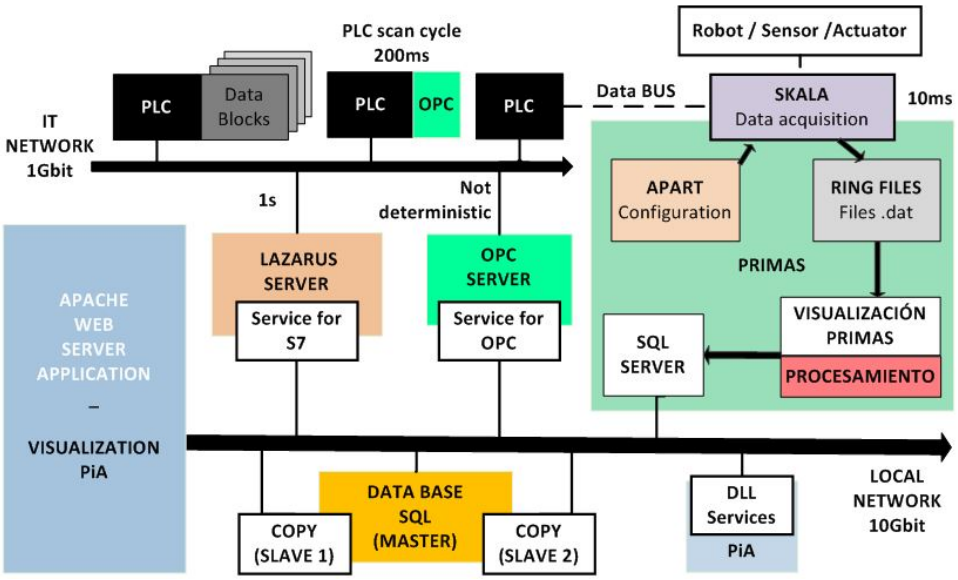


FIGURE 2.12: Data generation architecture in the Sindelfingen paint shop.

data source point associated, these are, from left to right, the data blocks in the PLC memory, the OPC connection and the communication bus with the actuator or sensor correspondent.

Between both networks, there are servers that support the LAZARUS and OPC data generation methods. These are connected to the data generation points of the equipment through the IT network. The services implemented in these servers deposit the data, through the connection to the LOCAL network, in the information management systems of the paint plant data equipment (DLL-PiA Services and the SQL database, of type PostgreSQL, which is the central repository for process data). Acquisition times are approximately 1 second for the Lazarus system (depends on the PLC cycle time) and is not time deterministic for the OPC system (there are no predetermined acquisition periods). This makes them not suitable solutions for generating data in real time. On the right side, there is the PRIMAS real-time data generation method, which dumps the data generated in the LOCAL network directly. This data, as indicated, has an acquisition period of 10ms, this allows data capture in real time.

Methods for data generation in the Sindelfingen paint shop

The different data generation methods used at the Sindelfingen paint plant are described below. The structure of the sections includes a theoretical framework of the method, application examples and, in the most relevant ones, a specific evaluation of the method for data generation.

- LEVEL 4 - PRIMAS

  PRIMAS is a proprietary system that allows the industrial data management. This system extracts data frames from the communication buses of the equipment of the paint shop (Profibus, SercosIII), translates these frames into usable data and makes them available to the factory data analysis team. PRIMAS allows data capture rates of 10 milliseconds so it works in real time.

– PRIMAS system architecture

The architecture of the PRIMAS system is presented in Figure 2.13; shows the integral management of data in real time, from the part of acquisition of raw data to the part of exporting data already usable to other systems of the plant.

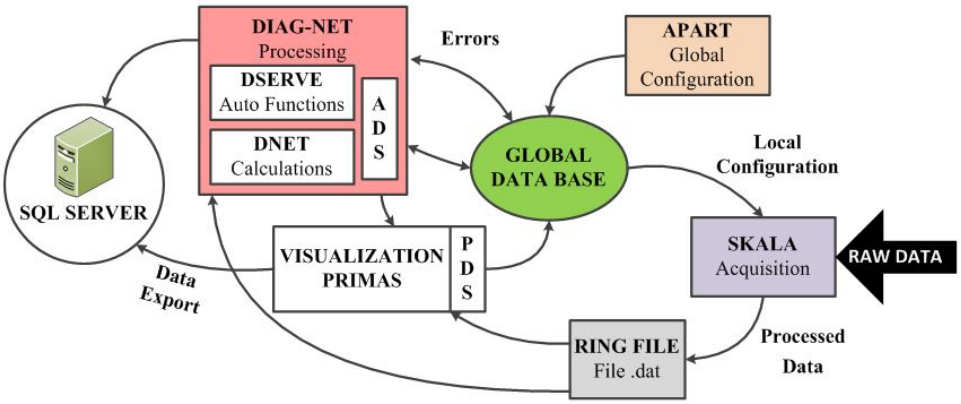


FIGURE 2.13: PRIMAS system architecture.

In Figure 2.13 the following main components are distinguished:

* APART: it is the global configuration tool of the system. All the necessary configuration parameters are generated for each case of data extraction and transformation of the communication buses. These parameters are stored in the system database, GLOBAL DATABASE and, from there, the local configurations for each particular data generation case are distributed.
* SKALA: this module is in charge of, once it receives the local configuration parameters, reading the bus frames, processing the raw data, translating them into usable data and sending them to the ring files.
* RING FILES: ring files are FIFO type (.dat) files that allow fast input and output of data. Each file takes care of the values of certain variables. Data from the SKALA modules enters and is distributed to visualization or processing.
* PRIMAS VISUALIZATION: it is a system that is in charge of data management and has its own PRIMAS interface in which plant workers can make inquiries about the status of the values of the variables. It does not allow data to be extracted to be used by Mercedes Benz applications, but it does export information to the SQL SERVER from which this information can be accessed by applications developed in the plant.
* PDS: "Prozess Daten Streaming" is a data cloud that forms an intermediate layer between the data sources (translated and usable data) and the algorithms or server processes (data visualization, processing).
* DIAG-NET: system data processing module. It is made up of DSERVE and DNET.

* DSERVE: service for the automation of the PRIMAS process. They are the set of functions that allow the system to operate automatically.
* DNET: it is a calculation service within the PRIMAS system that allows operations to be carried out with data and generates rules, error signals or new data.
* ADS: "Aggregierten Daten Streaming" is, analogous to PDS, a data cloud with the aggregated data that has been processed or generated in DIAG-NET.
* GLOBAL DATABASE: central database of the PRIMAS system, common data access point between system applications.
* SQL SERVER: it is a database of the PRIMAS system to which queries and data extractions can be made from the Mercedes Benz systems of the paint plant. It is the access point to real-time data from plant systems.

The starting point to consider in the PRIMAS architecture scheme is the global configuration tool, APART. This tool, an example of which is shown in Figure 2.14, is made up of a series of tables in which the necessary information is inserted so that the rest of the parts of the system can carry out their functions. This tool deposits the configurations in the GLOBAL DATABASE from which the rest of the systems will retrieve the corresponding local configuration.

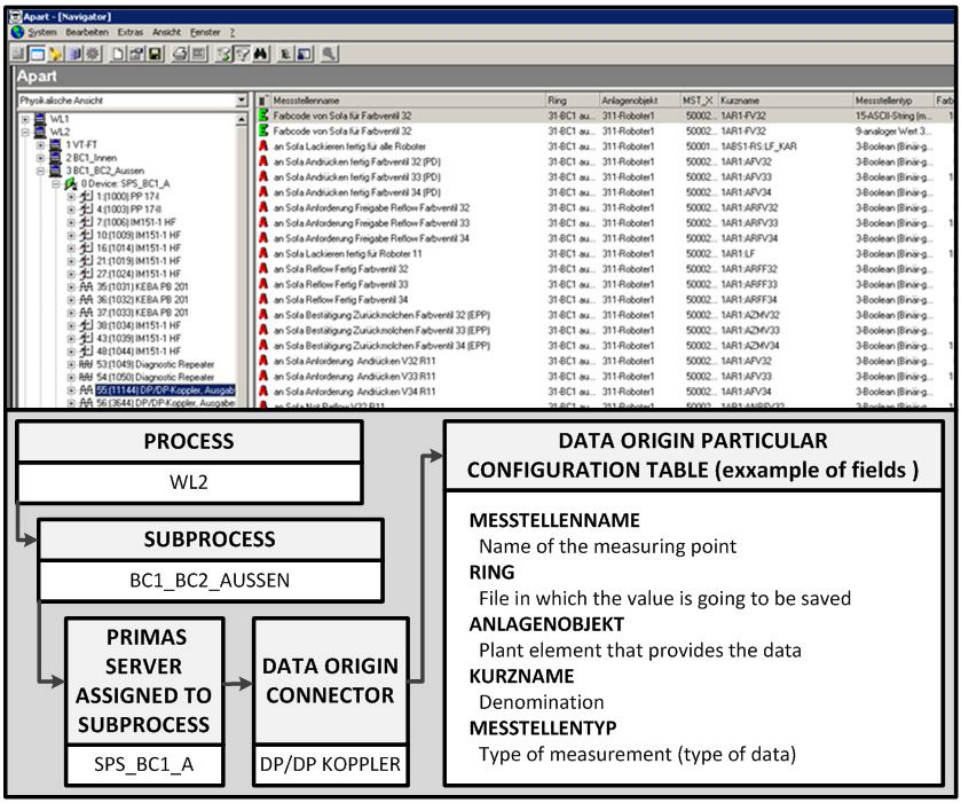


FIGURE 2.14: PRIMAS configuration tool: APART.

In this example, appears on the left, the division of configurations according to a physical view (Physikalische Ansicht) of plant data source devices that are classified by their location within the process as WL1 - WasserLack1, WL2 (BC1, BC2 ...) followed by the data acquisition server that is running in that part of the process and hence the connections to the data sources.

The next point to consider in the scheme presented in Figure 2.13 is data entry (RAW DATA). This data entry is done through a SKALA module. This module performs data acquisition through a connection to the corresponding data bus. A requirement in order to be able to extract data from the bus, without added costs for the process equipment, is that the connection is made passively, that is, the bus topology does not have to be altered. Connecting a new element to a bus that connects a machine and its controller and that has not been contemplated in the original programming can lead to errors on the bus. To avoid these errors, the elements of the bus would have to be reprogrammed, raising the cost of development and compromising the reliability of the system.

The solution to achieve this passive connection is shown in Figure 2.15. It consists of connecting the bus to a diagnostic module, such as the "SIMATIC Diagnostic Repeater for PROFIBUS-DP" since these modules have a mirror port in which they copy the signals that travel through the bus using optocouplers so as not to alter the topology of the bus. SKALA has interfaces for a multitude of field buses so it is capable of extracting data from any bus used in the paint plant facilities.

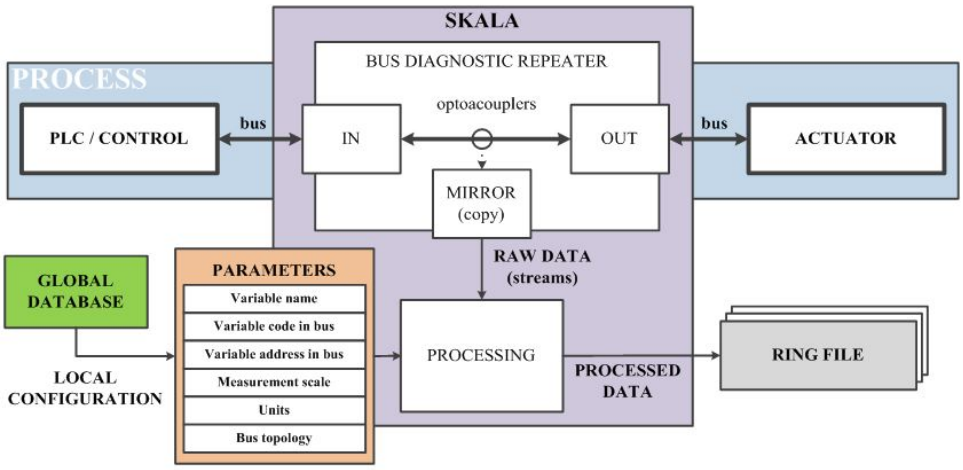


FIGURE 2.15: PRIMAS raw data acquisition tool: SKALA.

This data acquisition module is responsible for collecting the raw data frames of the bus and, using the specific local configuration tables of that data source, transcribes the format of each data within said frame and thus transform that original data into usable data. To carry out this transformation of the frames into usable data, it is necessary to know the following parameters for each frame that travels on the bus: name of the variable, variable code, address on the bus, scale, measurement units, bus topology. These data are necessary since each variable or signal value that

travels through the frame is encoded by the machine provider. The data frames that travel through the bus are made up of a group of parameters for each piece of data, which means that multiple decoding tasks have to be carried out each time. Being able to discover the correspondences between these parameters and their codes without the collaboration of the supplier companies is a limit in the generation of data. Having this information is one of the greatest added values in terms of data capture of the PRIMAS system. Once the raw data has been transformed into usable data, it is sent from the SKALA modules to the ring files. Ring files are ".dat" files (data files) in which the values of the variables are stored for a specified time. The file assigned to each variable is defined in the APART configuration table (RING field). From these ring files, the data is distributed to the PDS system of the PRIMAS system in the visualization section, or to the DIAG-NET processing. The PRIMAS system's own visualization system allows access to process data from any equipment in the painting plant with access to the local network. It is accessed through a connection to a remote server where this interface is installed and is dedicated to supporting this viewer. The functionalities and the interface are proprietary and the plant solution development teams do not have access to modify the functionalities of the system, the user can configure the data capture, save a record every few seconds or minutes, they can select the elements to be displayed and the display format, all within the possibilities offered by the system. The visualization can also show data generated by the processing part. The data that feed this visualization system comes from the PDS system. Services can request process data from this cloud over the network (data export in Figure 2.13). The PDS system interfaces between process data and data generated in the raw functions and visualization systems or external functions that require data from PRIMAS, such as PiA, which is the data visualization system developed by the Sindelfingen applications developer group in which data acquired through PRIMAS are complemented with data from other sources such as other generation methods, aggregated data, the production plan, quality data. Thanks to the PDS system, the volume of data in PRIMAS is reduced by requiring only one copy of the data, all data can be consulted in a synchronized way including calculated data (they are subsequently synchronized), the algorithms can independently write the process data (PDS handles multiple data inputs and multiple data outputs simultaneously), has the ability to share data in real time (data flow in the shortest possible time intervals), error detection (by including data processing, different levels of data loss and waiting times are detected and resolved), self-organization of data flows between nodes (generates an order in the local interaction between system components).

The processing part, DIAG-NET, is made up of DSERVE and DNET. DSERVE is made up of the functions that allow the PRIMAS system to operate automatically and DNET is the calculation system. These systems aggregate data, perform calculations on the process data and, if necessary, generate error signals through the result of those calculations and provide end-user process evaluations through defined rules of process operation. Both are developed by Techno-Step with the needs of the paint plant data team in mind. ADS, analogous to PDS, makes the data aggregated and generated

in the processing available to the system. Remembering that PRIMAS is a closed data management system, the SQL SERVER is the point of the system from which PRIMAS data is made available for other applications and plant services.

– Examples of data generation with PRIMAS

  * Data extraction from sensors and controllers not connected to a communication bus

    This first case shows the structure and architecture of data generation with the PRIMAS method used to extract data in real time from sensors that are not connected to a field bus.

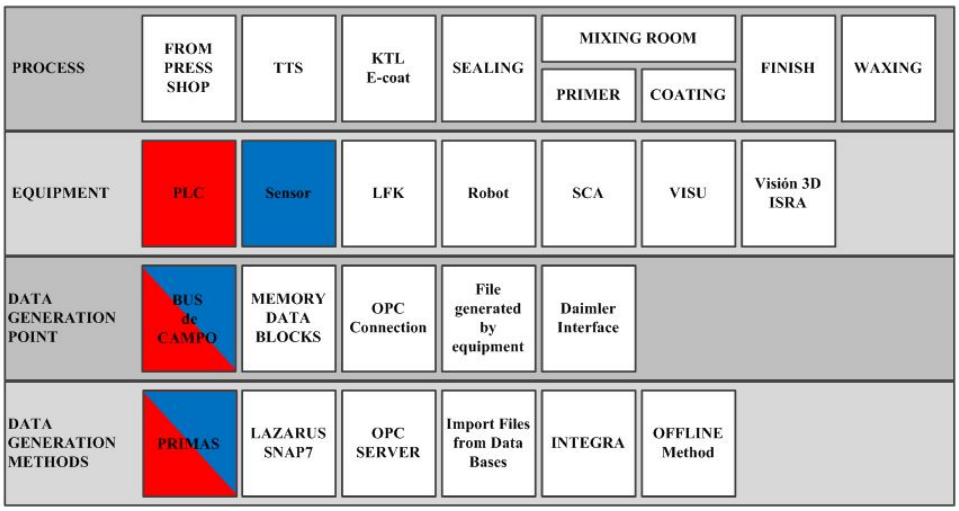


FIGURE 2.16: Generation structure for the PRIMAS Sensors case.

As indicated in Figure 2.16, the generation structure of this case presents the acquisition of data in real time by PRIMAS in two situations. In blue, the acquisition of data from independent sensors, without connection to the controller, used in the process is presented. In red, the extraction of data of interest from the controller that is not currently circulating on the fieldbus used by the PLC in the process control.

Figure 2.17 presents the architecture for the generation of data in real time in the cases of sensors not connected to a bus (left part of the image) and data of interest from the PLC that are not currently circulating on the field bus of the process (right part of the image).

On the left side, data acquisition is carried out through the connection of a PRIMAS SKALA module to a field bus of the PRIMAS system, Interbus, to which the independent sensors whose real-time data are of interest are connected. On the right side, a fieldbus is used between the controller and a ghost-ghost element. This element simulates the presence of an actuator and allows the controller to be tricked into sending the data of interest on the bus. These data are acquired by connecting a PRIMAS SKALA module to this bus.

  * DÜRR painting machine data extraction

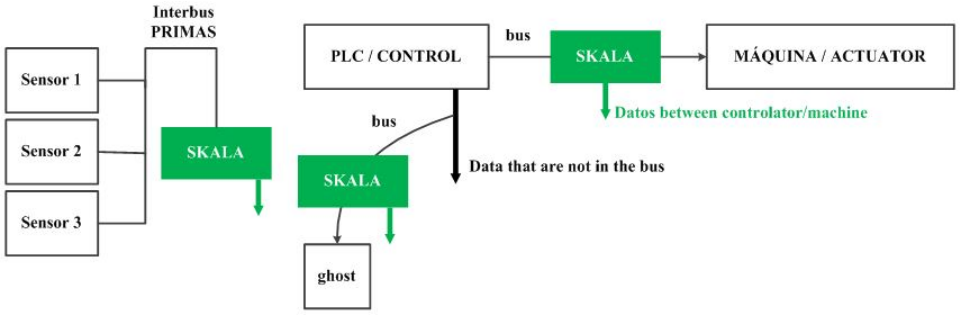


FIGURE 2.17: Extraction of data that is not on a fieldbus with PRIMAS.

As shown in Figure 2.18, the generation structure of this case presents the acquisition of data in real time using PRIMAS in two situations. In purple, the DÜRR Robot data acquisition (DÜRR machine) is presented. In red, data extraction from the DÜRR controller via fieldbus used in process control.-

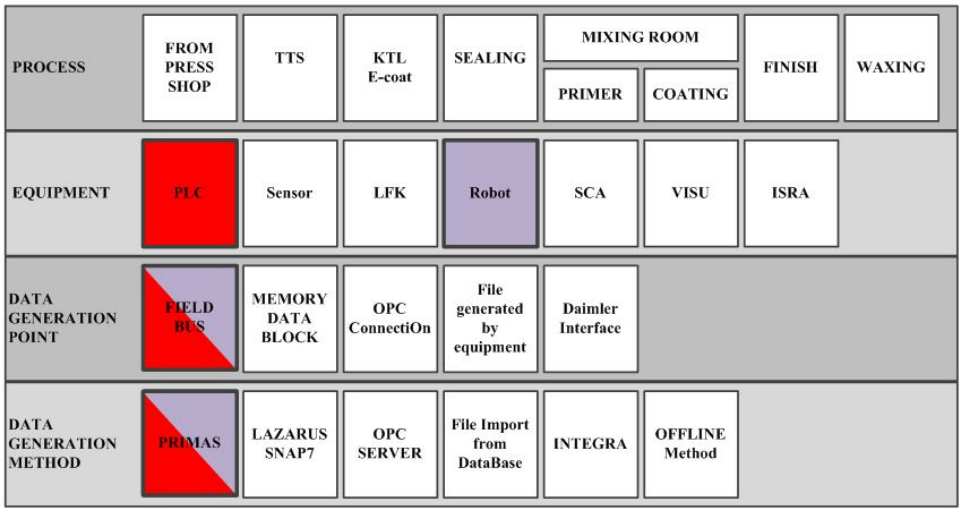


FIGURE 2.18: Generation structure for PRIMAS DÜRR case.

As shown in Figure 2.19, DÜRR painting machines can have two types of buses, Profibus (old machine) or Sercos III (modern machine). PRIMAS copies the raw data frames that travel through these data buses and through the configuration parameters it is able to convert this raw data into available and usable data. The configurations generated to translate the data frames are the result of collaboration agreements between Techno-Step and DÜRR. An added limit in the case of DÜRR machines is that parts of the control have components from other suppliers, such as Bosch in this case, and contain some data that cannot be accessed, such as the control of the robot axes. As shown in Figure 2.19, it is a "black box" within which there are several control systems from other providers that contribute data to the DÜRR control through DP / DP couplers, whose function is to enable the exchange of data between two bus owners (DP Master)

without there being a physical connection between both individual buses. This black box has a data storage system sent by these control systems that do not belong to DÜRR called Intradrive whose data cannot be accessed since they are the property of this second provider. This dependence on providers to obtain usable data poses a threat to the PRIMAS system since it is possible that data from any part of the machine that does not directly depend on a main provider can never be accessed.

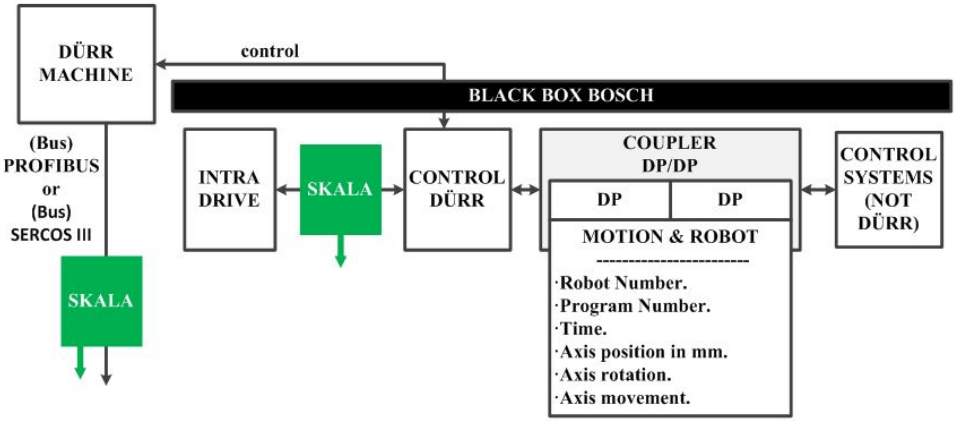


FIGURE 2.19: Data extraction in PRIMAS DÜRR case.

* SAM (SCA + Robot)

This case represents the structure and data generation architecture of the SAM sealantsellador application system. The SAM sealant application station combines a sealant pumping equipment, SCA, a Sealant Application Robot and a PLC as station control equipment.

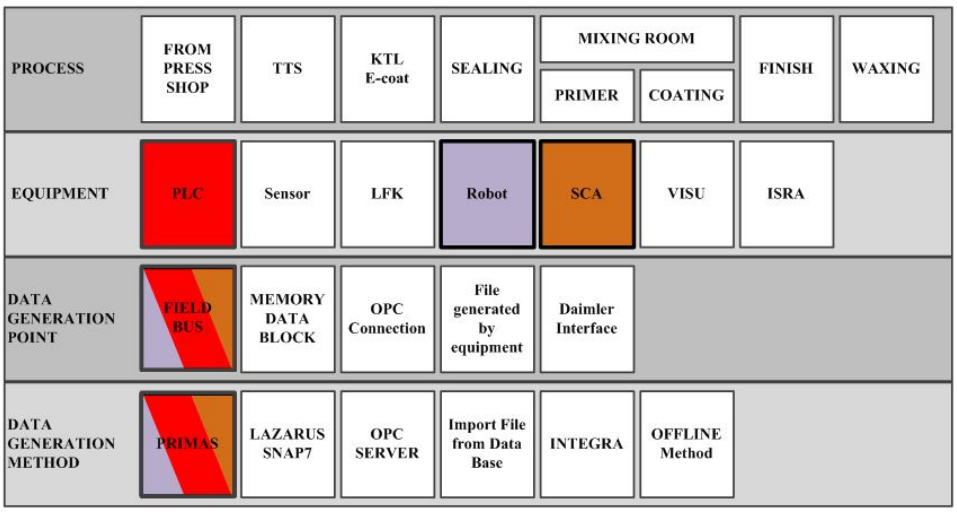


FIGURE 2.20: Generation structure for PRIMAS SAM case.

As shown in Figure 2.20, the generation structure of this case presents the acquisition of data in real time using PRIMAS in three situations. In red, the SAM station controller data extraction. In purple, the data

acquisition of the Robot used for the sealant application is presented. In brown, the data acquisition from the SAM SCA sealant pump system is depicted. Figure 2.21 presents the generation architecture of the case, with the generation of data through a connection of SKALA modules from PRIMAS to the field buses of the SCA pumping systems, Robot and station control, PLC.

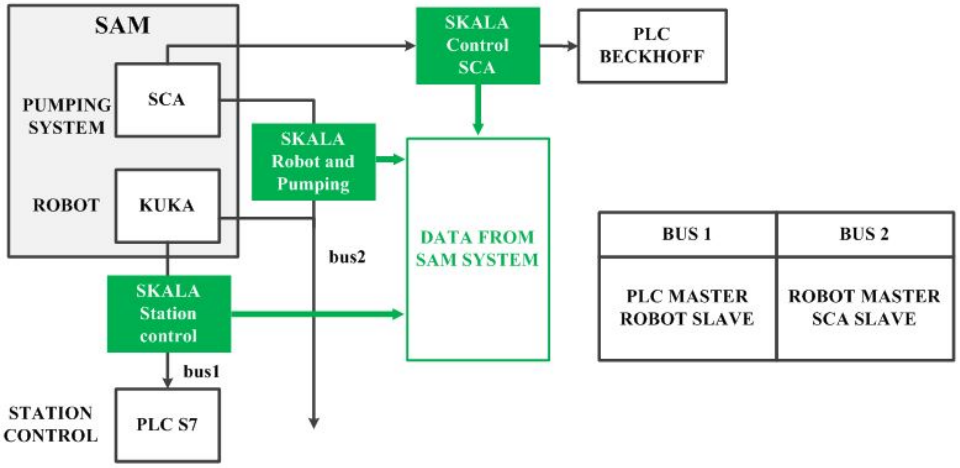


FIGURE 2.21: Data extraction in PRIMAS SAM case.

As shown in Figure 2.21, the PRIMAS system is capable of generating usable data from the data frames that are exchanged between the station control and the Application Robot; between the Robot and the SCA pumping system and between the SCA pumping system and its controller, in this case a Beckhoff brand PLC. As in the previous case, the configuration tables that allow the translation of the frames into usable data have been obtained and are the property of Techno-Step. In this case, bus1 is the one that relates the control of the station to the robot (MASTER of bus 1), among others, start and end of process signals are communicated to the robot (SLAVE of bus 1) and the robot communicates your ready signal to start the process. On bus 2, it is the robot (MASTER of bus 2) that indicates start signals to the application system (SLAVE of bus 2). Between the SCA system and the Beckhoff PLC, data is obtained on application pressures, application time, amounts of sealant applied.

* ISRA

This case represents the structure and data generation architecture of the case of a station that uses a Vision3D system, ISRA. The station consists of a Robot that is going to perform an operation on the bodywork, a PLC as the station's control equipment and an ISRA system to send the corrections to the Robot's trajectory according to the real position of the body with respect to the expected one. The generation structure of the case is presented in Figure 2.22.

As shown in Figure 2.22, the generation structure of this case presents the acquisition of data in real time using PRIMAS in three situations. In red, the extraction of data from the controller, PLC, of the station.

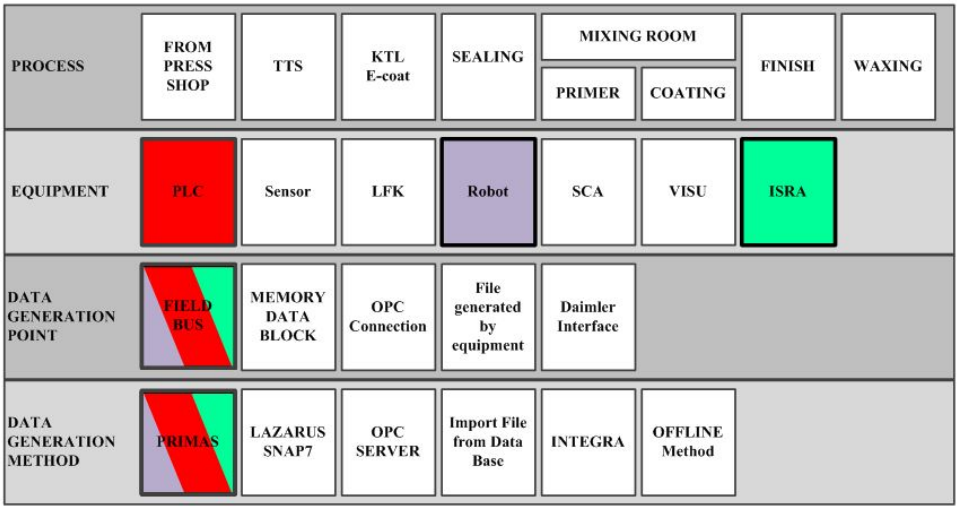


FIGURE 2.22: Generation structure for PRIMAS ISRA case.

In purple, the data acquisition of the Robot that is going to perform the operation on the body is presented. In turquoise green, the acquisition of data from the station control system to the ISRA system is represented. Figure 2.23 represents the data generation architecture through a connection to a PLC, a Robot and a 3D Vision system, ISRA, using SKALA modules of the PRIMAS system.

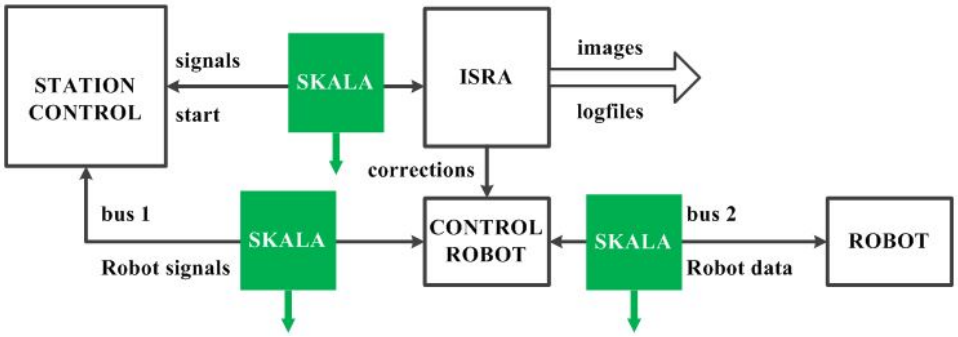


FIGURE 2.23: Data extraction in PRIMAS ISRA case.

The ISRA example, whose generation architecture is represented in Figure 2.22, is a Vision system using 3D cameras that measures the position of the body in the robotic stations, calculates the variations of the current position with respect to the expected one of the body and sends the corresponding corrections to the robot controller to adjust its trajectory. In this case, PRIMAS is used to capture the control signals of the equipment elements, that is, the station control PLC, receives the body positioning signal and sends a measurement start signal to the ISRA system. Signals from the controller to the Robot and data from the controller-robot bus are also captured. Images and log files, logfiles, are extracted using the file import method

– Analysis of the use of the PRIMAS system for data generation

PRIMAS is a proven real-time data generation system at the Sindelfingen paint shop. It is based on the extraction of data frames in the communication buses between controllers and actuators and has the ability to connect stand alone sensors to the data acquisition system. For all these reasons, it is a tool that is easily adapted to various data generation needs. Analysis of the PRIMAS method as a process data generation method:

* Strengths of using the PRIMAS system: the main competitive advantage of the system is, apart from its ability to operate in real time, having the tables that allow transforming the data frames extracted from the communication buses into usable data. Another strength is that it extracts the data frames without altering the topology of the buses, so it is not necessary to alter the control systems of the plant.
* Weaknesses in the use of the PRIMAS system: PRIMAS implies a series of limits for the company, such as that any modification of the configurations, functions, generation of rules, maintenance, hardware update, adoption of another operating system, go through the supplier company. This makes PRIMAS an expensive and relatively inflexible system, as well as functioning as a black box for paint plant data equipment. Also, being a comprehensive data management system, that is, it includes from the acquisition, translation, storage, processing, visualization and making available of the data to plant users, it has many critical points susceptible to failures, which It compromises its reliability and as maintenance is associated with the supplier, it may compromise the system indefinitely and for reasons that are not under the control and supervision of Daimler. PRIMAS also has a weakness of use at the beginning of a digital transformation of a plant. The complexity of the system means that process data is not immediately available, delaying the development of solutions using PRIMAS data to more advanced stages of digital transformation.
* Opportunities arising from the use of the PRIMAS system: some opportunities made possible by the use of the PRIMAS system is that, as it is a turnkey system for generating real-time data from industrial processes, it allows the plant's data team to focus on generating value with the data obtained. It also allows for a smaller and more specialized team of people. This makes it easier to develop process control solutions such as the digital process and product twin.
* Threats generated by the use of the PRIMAS system: the use of the PRIMAS system entails a series of risks, such as the change in the technologies of the machines. It also depends on the collaboration of the suppliers of the machine to be able to carry out the translation of the data frames, since performing the decoding blind is a job with an unfeasible cost.

With the arrival of paradigms such as Industry 4.0 in which data acquires great value, machinery manufacturers have realized that the data generated by their devices was an untapped value-added asset. This may mean that in future evolutions of these devices, the manufacturers themselves could provide their own data acquisition ports, with added cost, while

making it difficult to acquire data from their devices through other methods such as PRIMAS.

- LEVEL 4 - LAZARUS SNAP7

  It is a data generation method applied to the data blocks in the memory of Siemens S7 PLCs. These values of the data in the PLC memory are updated in each PLC cycle, this is every hundreds of milliseconds, so this method does not have the ability to generate data in real time. It is used for control signals, alarms and information generated by the PLC. LAZARUS SNAP7 refers to separate concepts that form a data generation method. SNAP7 is a free Ethernet communications suite (open source) that allows cross-platform applications (Node.js, .NET / Mono, Pascal, LabVIEW, Python and C / C ++) to be interconnected with Siemens S7 300 and 400 PLCs (partially with 1200 and 1500). Lazarus is a development environment, IDE, based on the Pascal programming language. It is the environment chosen in Sindelfingen for the development of communication functions.

  – How LAZARUS SNAP7 method works

    There is a LAZARUS server where the services are located. This server communicates, using the RPC1006 TCP / IP protocol, sequentially with the PLCs connected to the IT network, as indicated in Figure 2.24.

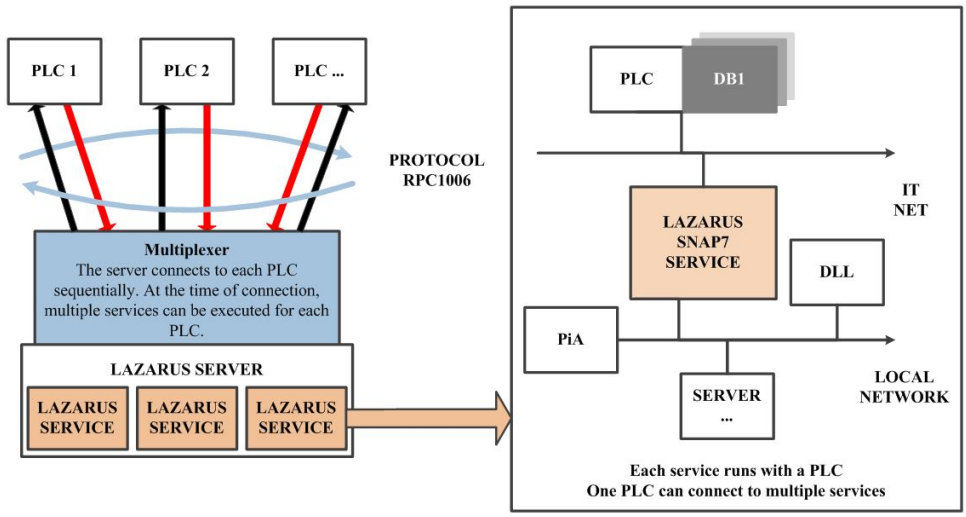


FIGURE 2.24: LAZARUS methodology.

The server changes the connection from one PLC to the next, and during the interconnection period, the services programmed to interact with said PLC come into operation. A service is only running on one PLC at a time, but a PLC can be connected to different services simultaneously. The services access the PLC memory through a connection provided by the SNAP7 library functions. The LAZARUS server can send data generated by the functions to the local network to be processed, visualized or stored by other components connected to that local network, such as the SQL server or the visualization of the PiA system.

– Data generation example with LAZARUS SNAP7

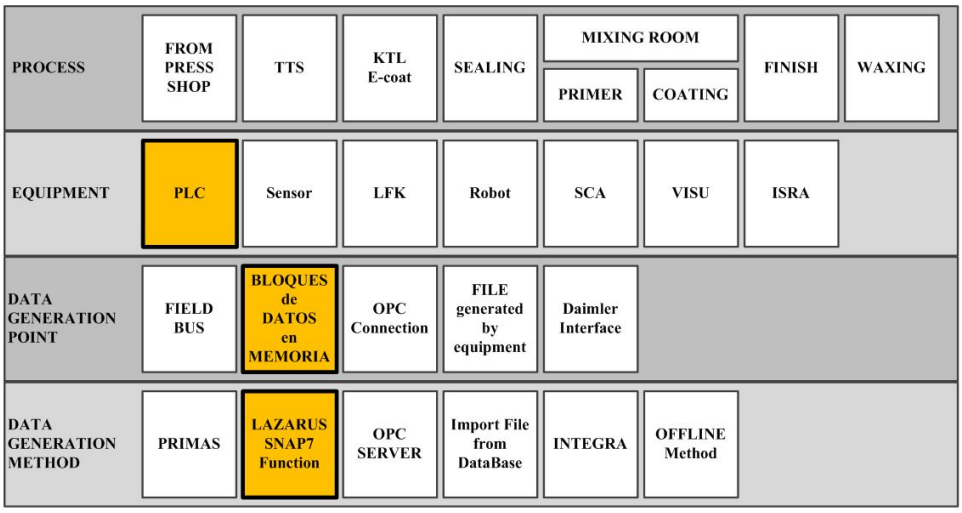


FIGURE 2.25: LAZARUS data generation structure.

As can be seen in Figure 2.25, using the LAZARUS STEP7 method, data is obtained that is stored in data blocks in the PLC memory. The example in Figure 2.26 represents a case of data generation using a LAZARUS SNAP7 function that is capable of reading data present in the data blocks in the PLC memory. The example service is created with the LAZARUS application development environment using the SNAP7 library. The objective of this service is to extract a piece of data from a given block of data from the PLC memory when it detects that there is a new value.

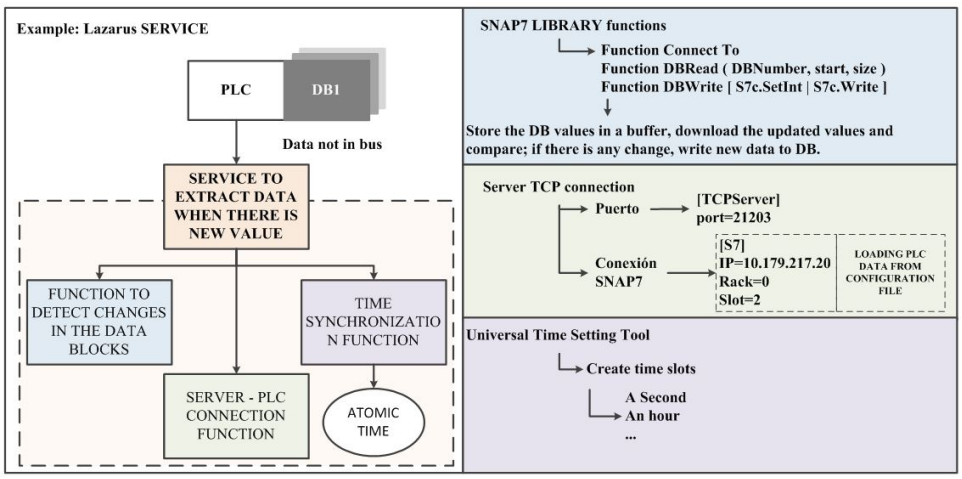


FIGURE 2.26: LAZARUS data generation example.

This service, described in Figure 2.26, is broken down into three functions: the time setting function establishes the time interval between checks of the data value (in purple). The time base provided by a global time synchronization tool is used that assigns a common time to all the equipment in the plant. The function to make the TCP Server - PLC connection, using the SNAP7 library (in green). Server and PLC are configured. In the case of the data to configure the PLC, it is loaded from a configuration file, so

that when updating any of the fields, it is modified in a single place. The function of detecting changes in data blocks. This function uses instructions available in the SNAP7 library to read and write to the PLC data blocks (in blue). There are values of a block of data stored in a buffer. The function connects to the PLC, reads the current values from said data block, compares the values between the stored and updated data and, if there is a change, updates the value. This service runs on the LAZARUS Server, as defined in the data architecture section.

– Analysis of the use LAZARUS SNAP7 method for data generation

In this analysis, two parts can be distinguished, one referring to the use of the LAZARUS development platform and the other referring to the SNAP7 library.

* LAZARUS: the strengths of LAZARUS are, that it has proven itself as a viable development environment to develop the necessary applications at the Sindelfingen paint plant and that it is a free development environment. A great weakness is that it is based on the Pascal language. This language is not at the top of the list of most popular languages (of those compatible with SNAP7, the most popular are Python and C / C ++), so it can be more difficult to find professionals who have experience in developing applications in this language and can also make it more complex to find support in the community (being free software, there is no official support). The opportunity of using LAZARUS is that, in a plant digitization process, it allows a relatively fast and inexpensive way, being free software, to start obtaining process data. The threats that derive from the use of LAZARUS are that it requires personnel with experience in Pascal language and support in terms of providing system security.

* SNAP7: the strength of SNAP7 is that it is a free multiplatform library that allows access to the memory of Siemens PLCs. It can be used in a variety of programming languages, including some of the most popular today. The weakness of using SNAP7 is that it is only compatible with Siemens S7 300/400 PLCs and partially with the more modern 1200/1500. This makes it not a comprehensive plant PLC data extraction solution. Another weakness is that it does not allow real-time data to be obtained. The opportunity arisen from the use of SNAP7 is that it allows having PLC data relatively early during the digitization of a process. One threat of using SNAP7 is security, it is open source and it is not common for it to receive support regarding vulnerabilities from the library.

- LEVEL 4 - OPC SERVER

OPC is a widely implemented standard that offers the ability to exchange data between industrial systems via an Ethernet connection. OPC uses a Client-Server approach for exchanging information in a bidirectional way. The OPC Server offers easy interoperability as it eliminates the need for custom-made interfaces or drivers as it communicates data sources with an OPC Client using its native protocol directly (there are OPC Servers for different brands of devices such as Siemens, ABB. The OPC Client can connect to the OPC Server to

read and consume information (alarm signals, events) that you want to transmit to the OPC Client.

- LEVEL 4 - File Importation, from equipment to DataBase

  This section refers to the cases of file importation by services located on the PiA Server from LFK, VISU and ISRA computers. These PiA Services read, transfer, process and store the data in those files continuously. The following describes the application cases of the file import method at the Sindelfingen paint plant.

  – In the first case, this import refers to the xml file generated by the paint error control installation, LFK. There are 3 LFK units connected to a central service implanted in the PiA Server. All raw data from the LFK results contained in the file is stored on the SQL server.

  – In a second case, an import is made, from a PiA server service, of a csv file generated in a VISU of the KTL process. This file contains SKID data (related to the vehicle's MDS number), water measurement values, KTL temperatures among others. These values are stored on the SQL server.

  – In the third case, there is a connection to a central service of the PiA Server to extract the files generated by ISRA Artificial Vision equipment. The log files generated and the images captured by the ISRA system are continuously read, transferred, processed and stored. The measurement results and logs are stored on the SQL server. The ISRA images are stored on a special ring server developed on purpose by the Sindelfingen data team.

- LEVEL 4 - INTEGRA

  It is a Daimler standard for hardware and software technologies used in most Daimler plants around the world. In the case of Sindelfingen, it is used as the data generation method of the VISU equipment, the display screens of the PLCs, through a standard library (Daimler interface). It is a very slow generation method, so it is only used to capture signals that are not needed in real time.

- LEVEL 4 - OFFLINE Methods

  Once the data generation methods implemented in Sindelfingen have been presented, it is worth mentioning some methods for obtaining machine data that, although they do not allow the development of a real-time data management system such as the one in operation in Sindelfingen, allows to obtain data that can lead to the development of off line solutions that help decision-making, such as analysing the most relevant variables for a specific aspect of the process. The validity of these methods is justified by the time and resources necessary to implement a comprehensive real-time data management system as presented in the previous sections on generation architecture and generation methods.

  – Siemens S7 PLC: To extract data from the PLCs, there is a software called generically "PLC Analyzer" that allows finding signals in Siemens devices. For example, the "SPS-Analyzer" software from Delta Logic has been recommended by the Sindelfingen team to carry out this task.

  - KUKA robots: These robots have, from the factory, an "Osciloscope Mode" that allows you to view data from the bus, but only within the system, so it is not valid to develop a solution that works in real time.
  - SCA: These mastic pumping systems have a Beckhoff controller. This controller, as in the previous case, has an oscilloscope mode that allows monitoring a signal indicating the start and end times of the measurement. This oscilloscope is called "TwinCAT Scope View" and is software available as standard on all SCA equipment.

Data synchronization methods in the Sindelfingen paint shop

In Sindelfingen it is possible to relate the data from various perspectives both on a time basis and on a process or product basis, that is, the values of the process variables can be obtained according to a date / time, a specific process or a determined vehicle body or type of vehicle body.

- Time synchronization of all factory systems

  The data systems of the Sindelfingen plant collect the time from the same atomic clock (atomic time). This allows exact time synchronization between all systems. All systems and services, both those of Daimler (PiA server) and those of the suppliers operating in the plant, use the same time provided by the time server. This distributes an update of the time every 10 milliseconds through an antenna to all other systems in the plant. Programs and functions developed for the generation of process data use the same time origin.

- Spatial synchronization, location of the body in the process

  When a new bodywork arrives at the paint shop, it has a locating device on which its MDS number is engraved, the unique identification of the chassis throughout the process. This allows a much simpler and more precise data synchronization, since the real-time location of the vehicle chassis within the plant is available. Each subprocess is divided into small body position feeds. This exact location of each body within the process is a key requirement to be able to synchronize the data using the body as the primary key of the dataset. It allows a synchronization of the values of the variables identified in that position of the process for each bodywork.

- Synchronization application examples
  - Process Digital Twin: The data of the relevant variables for each part of the process are synchronized to represent the operation of each activity carried out during the process. The times in which each variable is relevant to that activity are defined and its values are monitored during that time. In this way, rules can be generated that supervise the correct operation of said process. Knowing when the variables are relevant, it is possible to know which are the variable values that indicate a correct process and measures can be developed to monitor and guarantee this proper functioning of the process, such as creating alarm signals from the system's own servers. that supervise this digital twin of the process and send these signals through OPC protocol to the PLCs so that they can make the appropriate corrections or notify the anomaly to the plant personnel.

  - Chassis Digital Twin: at the Sindelfingen paint shop, this chassis digital twin is a file containing the values of more than 24,000 variables for each bodywork. These variables are selected by expert knowledge and represent the life of the body within the paint plant. The file is generated in the SQL database using trigger functions from the other databases. Due to data aggregation speed needs, arising in the creation of this file, the data is saved in JSON-B format since it is the dynamic way of storing the data that has given the best results and thus avoids excessively collapsing the databases of data. In this file, digital variables appear, such as presence in a certain place (TRUE / FALSE), values of analogical variables that are represented by their maximum and minimum value reached during the process (as explained in the previous point) and other categorical variables, such as the body model.

2.1.3 Process data analysis

Traditionally, data analysis systems have been designed to answer questions about the past; questions like, what has been the percentage of vehicles with defects in the paint film in the last month?. This was done by reviewing a group of process and quality control data stored on disks or cards, with high read times and was analyzed by processors with low computing powers. The results, some times hard to understand, had to be interpreted by expert knowledge, that is, professionals in the painting process with extensive experience. These interpretations helped company managers to make decisions, but they did not allow them to analyse the valuable information within the process data. Now, data analysis systems can answer questions from the future like: according to process data, will there be defects in the paint film layer of the vehicles?, or even, which process parameters must be modified to avoid defects in the paint film?. This analysis is based on the application of Data Analytics and Business Intelligence through the development of an MPC (predictive control model). This MPC is composed of a previous data acquisition stage and data quality evaluation stage as mentioned before.

Data analysis is defined as the process of drawing conclusions based on raw, untreated information. Through analysis, meaningless perceived data can be transformed into something valuable and usable. There are three subtypes of data analysis:

- Descriptive Analytics: it is a preliminary stage of data processing that creates a summary of the historical data to provide useful information and, in this way, prepare the data for further analysis.
- Predictive Analytics: analyses actual current and historical data to make predictions about the future or unknown events. Predictions cannot be 100% accurate, but they provide insight into what is most likely to happen next. This often involves data mining, machine learning, and statistics.
- Prescriptive Analytics: having a solid prediction of the future, a course of action can be prescribed. This turns data into action and leads to decisions in the real world. This section both data sciences involved, techniques and tools:
  - Data science (specific to the scope of the thesis):
    * Statistics.

    * Machine learning: tries to create programs capable of generalizing behaviours from information provided in the form of examples.
    * Data mining: tries to discover patterns in large volumes of data sets.
    * Artificial intelligence: a machine imitates the human cognitive functions, such as, learning and problem solving.
    * Operations research: is the application of scientific principles to business administration, providing a quantitative basis for making complex decisions.

- Techniques:
    * Linear regression: a mathematical model used to approximate the dependency relationship between a dependent variable Y, the independent variables and a random term.
    * Non linear regression: another mathematical model with which it is intended to obtain the values of the parameters associated with the best fit curve.
    * Logistic regression: is a type of regression analysis used to predict the outcome of a categorical variable based on the independent or predictor variables.
    * Time series model: which is a type of statistical inference that is made about the future of some variable or compound of variables based on past events.
    * Optimization: which is a method to determine the values of the variables involved in a process or system so that the result is the best possible.
    * Test A / B: a term used to describe randomized experiments with two variants, A and B, one being the control and the other the variant. The objective is to identify the changes that increase or maximize a certain result
    * Clustering (clustering algorithm): is a procedure for grouping a series of vectors according to a criterion, usually distance or similarity.
    * Factor analysis: is a statistical data reduction technique used to explain the correlations between observed variables in terms of a smaller number of unobserved variables called factors.
    * Principal component analysis: is a technique used to reduce the dimensionality of a data set.
    * Neural networks: a series of techniques that make it possible to find the combination of parameters that best suits a certain problem.
    * Support vector machines: a set of supervised learning algorithms used to build a model that predicts the class of a new sample.
    * Bayesian techniques: a subset of the field of statistics in which the evidence about the true state of the world is expressed in terms of degrees of belief or, more specifically, Bayesian probabilities.

    - Survival analysis: is an inferential technique whose main objective is to model the time it takes for a certain event to occur.

  - Tools:

    - R, SAS, Python, Java, C++, SPSS, MATLAB, Minitab, CPLEX, GAMS, Gauss, Tableau, Spotfire, VBA, Excel, Javascript, Perl, PHP, MySQL, AWS, Online solutions.

2.2 Critical study of the state of the art

When conducting the critical analysis of the state of the art, the first thing to mention is that in the literature there are no examples of the use of industrial process variables to develop predictive or prescriptive analytic applications to generate value. In other areas, such as banking, data sets related to them are studied on a regular basis for, for example, predicting a customer's churn.

Nor is there a clear methodology on how to generate a valid data set for the development of these advanced analytic applications. In fact, many methodologies are based on generating a messy data lake.

Industrial processes are very varied and complex. It is possible that many of the data extraction methodologies applicable in an industry are not interchangeable with others, this is due to, among other reasons, the different nature of the variables of each process.

In industry, many of the processes are governed by the expert knowledge of plant workers. There is no clear idea of which combination of variable values determines whether the result of the process is of the correct quality. In fact, many of the measures to ensure quality are corrective, leading to the generation of waste and superfluous expenses. Could the process data be used to predict the outcome in terms of process quality?.

2.3 Hipothesis

Considering the research gaps observed in the critical review of the state of the art, the following hypothesis is presented:

> An automotive paint shop production process can be modelled as a set of variables with linear and non-linear relationships in the form of a predictive model to estimate the body paint quality with a certain level of confidence and using it as a method to apply predictive control to the process.

2.4 Objectives

Objective 1: analyse the state of extraction of variables from the factory..

Objective 2: define a set of output variables that can best describe quality and that can be measured.

Objective 3: for this or these output variables, generate a set of input variables that can explain the outcome of the output variables.

Objective 4: describe a methodology to develop a predictive model of the paint film quality of the car bodies.

Objective 5: apply and test these methodologies in a real paint shop.

CHAPTER 3

Data set definition

3.1 Introduction

The objective of the definition is the univocal identification of a set of variables that characterize the process (input variables) and a series of representative indicators of the quality of the process (output variables). This process quality evaluation must be carried out as close as possible to the respective process, that is, it is convenient that each sub-process should have a measurable output variable that shows the process performance at all times.

In order to achieve this, the data set definition is divided into three steps: the characterization of the process, the definition of variables and the univocal identification of variables.

3.2 Process characterization

First, the characterization of the process is carried out, that is, going through the process so that the work flow is analysed and divided into a sequence of sub processes.

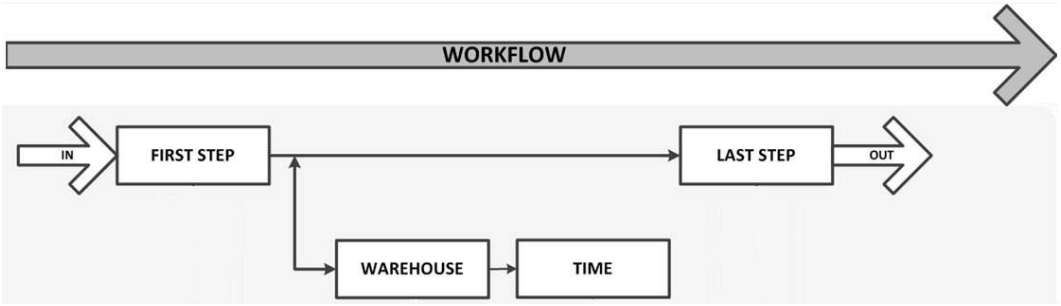
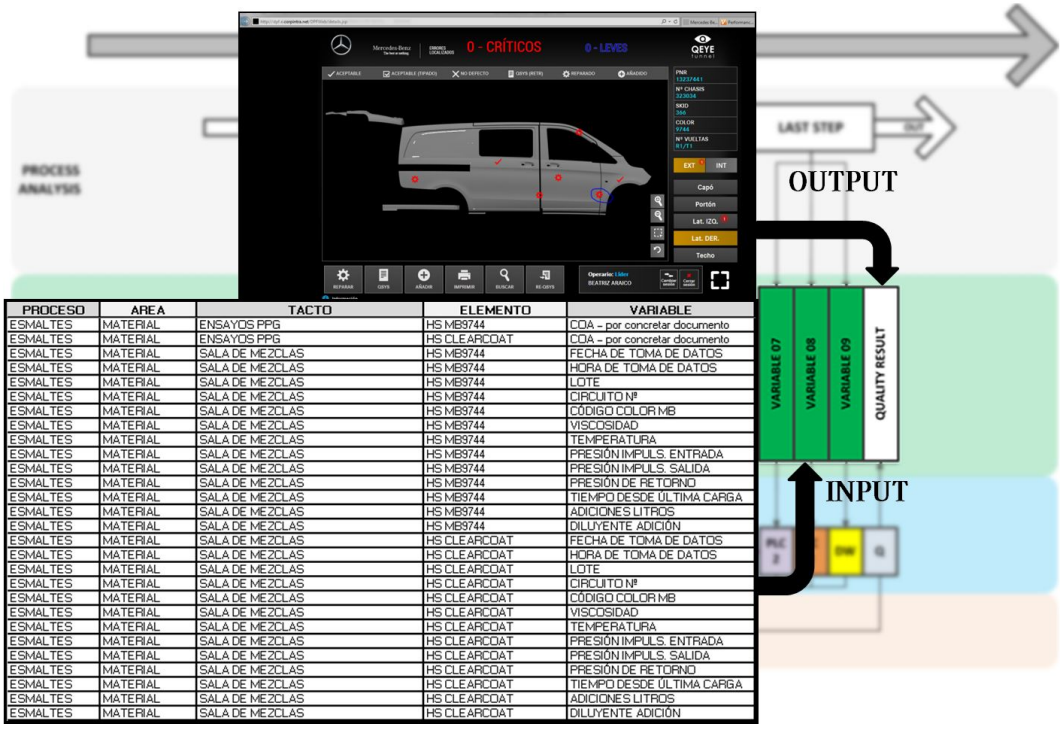


FIGURE 3.1: Process characterization procedure.

An application example is presented in Figure 3.1.

3.3 Variable definition

In the variable definition, in each step of the work flow, origins of variables are searched, such as, process control sheets, lists of sensors installed that in the area, control processes. That list of variables can include, input variables, output variables or disturbances that affect that section of the work flow. Once the variables are identified, they are assigned to that part of the process, as it is is presented in Figure 3.2.



PROCESO	AREA	TACTO	ELEMENTO	VARIABLE
ESMALTES	MATERIAL	ENSAYOS PPG	HS MB9744	COA – por concretar documento
ESMALTES	MATERIAL	ENSAYOS PPG	HS CLEARCOAT	COA – por concretar documento
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS MB9744	FECHA DE TOMA DE DATOS
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS MB9744	HORA DE TOMA DE DATOS
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS MB9744	LOTE
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS MB9744	CIRCUITO Nº
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS MB9744	CÓDIGO COLOR MB
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS MB9744	VISCOSIDAD
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS MB9744	TEMPERATURA
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS MB9744	PRESIÓN IMPULS. ENTRADA
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS MB9744	PRESIÓN IMPULS. SALIDA
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS MB9744	PRESIÓN DE RETORNO
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS MB9744	TIEMPO DESDE ÚLTIMA CARGA
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS MB9744	ADICIONES LITROS
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS MB9744	DILUYENTE ADICIÓN
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS CLEARCOAT	FECHA DE TOMA DE DATOS
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS CLEARCOAT	HORA DE TOMA DE DATOS
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS CLEARCOAT	LOTE
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS CLEARCOAT	CIRCUITO Nº
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS CLEARCOAT	CÓDIGO COLOR MB
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS CLEARCOAT	VISCOSIDAD
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS CLEARCOAT	TEMPERATURA
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS CLEARCOAT	PRESIÓN IMPULS. ENTRADA
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS CLEARCOAT	PRESIÓN IMPULS. SALIDA
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS CLEARCOAT	PRESIÓN DE RETORNO
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS CLEARCOAT	TIEMPO DESDE ÚLTIMA CARGA
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS CLEARCOAT	ADICIONES LITROS
ESMALTES	MATERIAL	SALA DE MEZCLAS	HS CLEARCOAT	DILUYENTE ADICIÓN

FIGURE 3.2: Origins of the identified variables.

3.4 Variable univocal identification

In companies of a certain size, such as, a vehicle assembly factory, with complex processes and extensive equipment, multiple skills and knowledge are required by their workers to manage knowledge within the organization, for this purpose Work is distributed based on specific skills.

regarding facilities and machinery, knowledge about the operation of equipment at the process level (aspects that affect production), electricity and mechanics (functional aspects), telematics (related to information and communications) can be differentiated. Therefore, it can be said that for the same system, different specialized teams are involved, contributing with their expertise and extracting their needs from each case.

These needs make them develop their own methodologies and applications, affecting only their scope. The problem can occur when, for example, in the case of developing an advanced analytic solution, as the equipment must be analysed from multiple points of view, since all can influence the generation of the said solution.

Then, when studying the know-how of each group, it conflicts with the multiple methodologies and forms of work of each one of them.

This is exemplified in this section, by the nomenclature of the same variable, considered from different points of view and needs of each group. This is reflected in Figure 3.3.

The production team is interested in controlling the stability of the process, so it is normal that its priority is the characteristics of the machine related to the process and define them by reference to the equipment (PUMP P1 PRESSURE in Figure 3.3)in which the variable is located. The maintenance team is interested in the functional aspect, related to the point of the controller (controller variable address in Figure 3.3 where the sensor signal that measures the characteristic of interest, pressure, arrives. The IT team is concerned about the value of the signal as it is registered in the server's memory (Server data address in Figure 3.3), with the name of this register representing the signal.

Data identification and data generation is difficult as these different variable denominations affects the communications between the different groups and the analytic solution generator, as the same variable is referred with three different names within the same factory. A identifier harmonization of the variable, so as the same variable code is know by all the work groups of the factory should be carried out, so that they have their own denomination that does not alter their methodology but they have another that provides work capacity with transversality.

There are multiple coding strategies, especially known as good practices using a specific programming language. The readability of the new codification, the non-compatibility of some symbols or terms because they are reserved words or that the codification makes any sense to anyone who meets that denomination.

3.4.1 Variable identifier harmonization

The same variable can have various names in the same Factory, as it is considered a process variable (process control), a measure signal (data acquisition systems) or an stored data (naming from databases).

An application example is presented in Figure 3.3.

3.4.2 Generation of an univocal variable codification

Once the list of variables identified in each step of the process has been obtained, a unique reference has been generated for that variable, so that it can be easily located within the process and encompasses the different names that each department has for that variable. The procedure followed for this unique identification consists of assembling a numerical code determined by the environment in which the variable is included. This environment is composed of a series of levels whose definition is determined by the adequacy to all the precision needs of each organization, that is, the number of levels that define a variable (process, sub-process, equipment, element, variable).

An application example is presented in Figure 3.4.

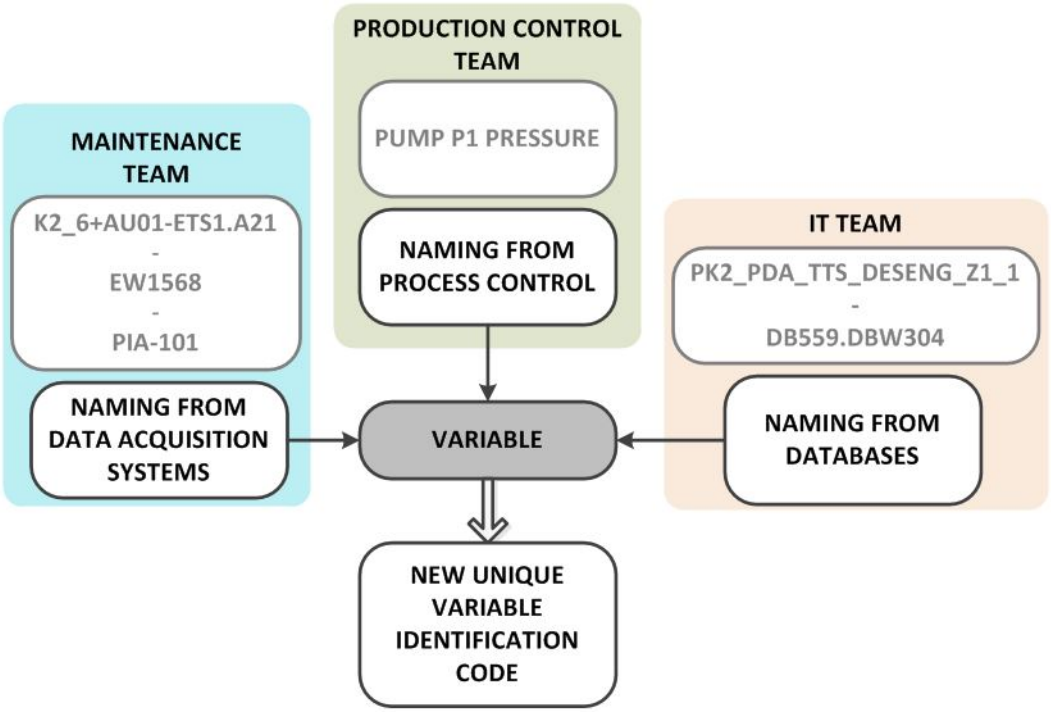


FIGURE 3.3: Variable identifier harmonization.

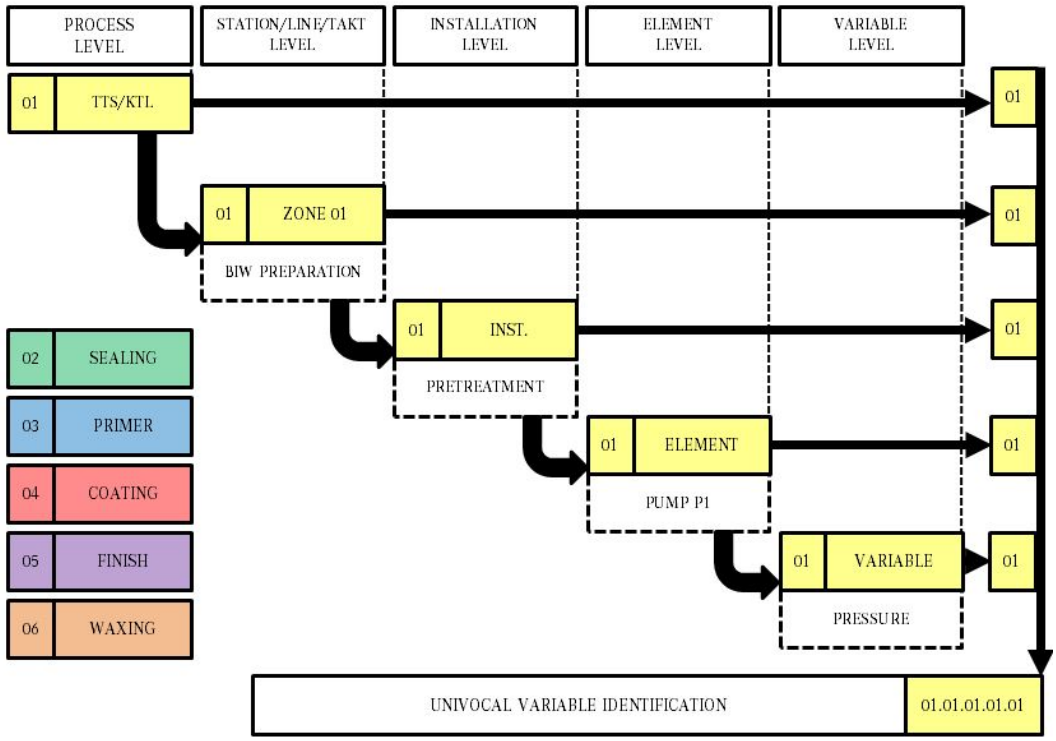


FIGURE 3.4: Generation of an univocal variable codification.

CHAPTER 4

Data set generation

4.1 Introduction

The objective of the generation part is to obtain the data set that will be used for the development of the advanced analytic solution, thus, it has been divided into two steps:

- Determine the origin of the data corresponding to the variables identified in the data definition procedure.
- Determine the data extraction methodology for each identified data origin.
- Synchronize this extracted data to generate a large table that feeds the analytic solution algorithm.

4.1.1 Determine the origin of the data for each variable

Once the list of identified variables has been generated, an analysis of the origins of these variables is carried out, identifying several origins as shown in Figure 4.1. Previously, a first classification of variables is carried out with the help of expert knowledge, determining if they can be considered relevant for the process outcome. This relevant categorization is performed in order to save efforts on the development of the project, reducing the presence of variables that only contribute to the generation of noise to the model.

According to these origins, the variable is classified as follows:

- The variable is not measured or cannot be measured.
- The variable is measured with a non-digitized instrument.
- The variable measured value is recorded in a non-digital format.
- The variable is measured in the process equipment and its value may or may not be stored in the machine's memory.
- The variable is used for control or display in a PLC and its value may or may not be stored in the PLC.
- The variable is registered in a database and its value is accessible.

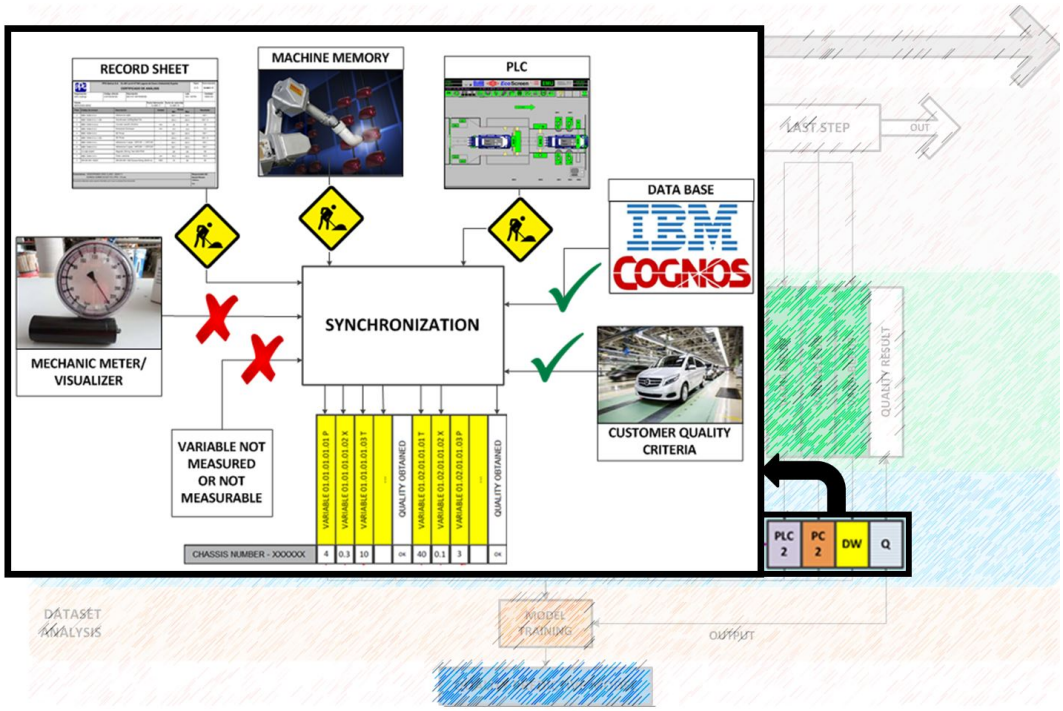


FIGURE 4.1: Identified variable data origin.

Determine the data extraction methodology for each identified data origin

This variable classification presents a data generation limit regarding the data generation methods. It is not always going to be possible to extract data directly from a desired variable. In this case, other methodologies can be tested, for example, using a proxy variable or developing a predictive model in order to estimate the value of that variable.

This is one of the greatest problems that industrial companies have when they develop advanced analytic solutions: they are not able to extract all the process data, so it will have a cost in the performance of the developed solution.

4.1.2 Data synchronization

The second relevant issue in generation, in addition to data extraction methods, is the synchronization of process data. The USB concept is proposed as a synchronization method.

The USB concept is presented in Figure 4.2. This USB concept, referred to a manufacturing industry, builds the data set by forming a table in which the rows are unique identifiers of each manufactured product (for example, the chassis id in automotive industry) and the columns of that table are the variables identified in the manufacturing process.

The data sources offer measurements of the process variables continuously but the table will only show the values of the variables when they were relevant to affect the KPI that is going to be analysed in the analytical solution to be developed. For example, in the case of wanting to determine the quality of the paint layer of a chassis

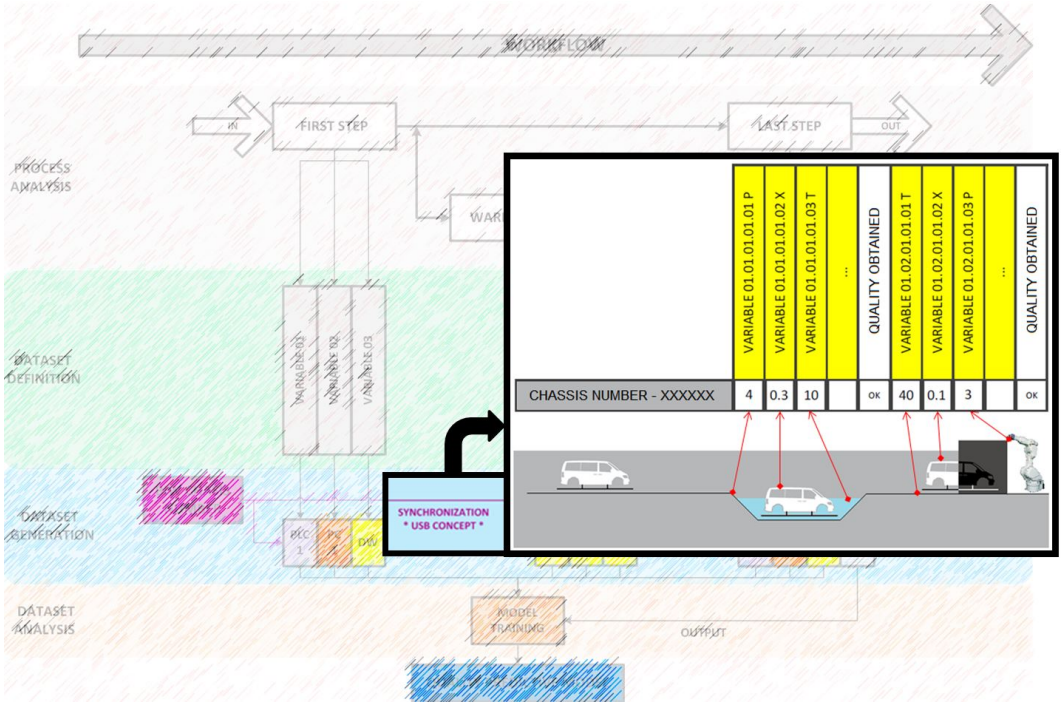


FIGURE 4.2: USB concept for data synchronization.

that circulates through the paint shop, these variables were relevant at the moment in which they "touched" the painted surface of the vehicle.

This synchronization has a series of considerations to take into account in order to be carried out satisfactorily. All the plant's data generation machines have to work in the same Timestamp, so that the temporal ordering of the sequence of actions that make up the manufacturing process is possible. The situation of the manufactured product within the process must be known precisely and continuously, so that the measured values of the variables can be assigned precisely at the moment in which the product was affected by said variables.

CHAPTER 5

Data set analysis

5.1 Introduction

The objective of this part is to develop the predictive model, it is presented in Figure 5.1. It has been divided into the following parts: collecting this data, analysing the quality of this data, developing prototype models, selecting the most appropriate model, and validating and deploying the model.

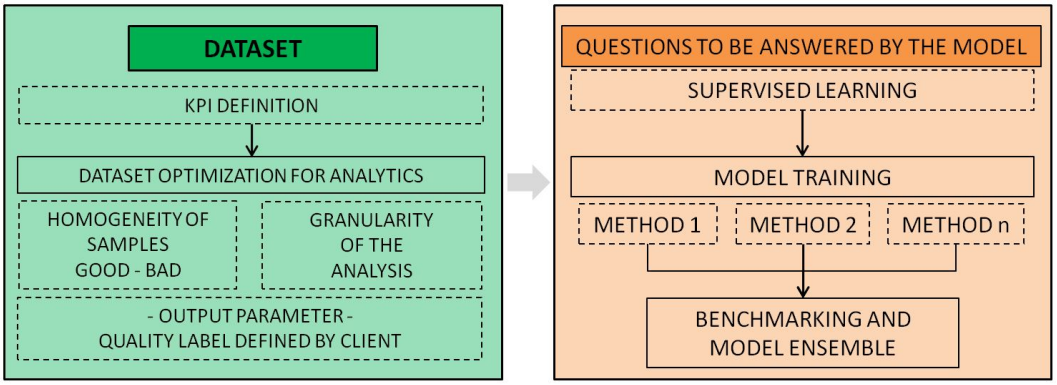


FIGURE 5.1: Data set analysis and modelling.

5.2 Data Collection

Once the variables that are relevant to the process whose output is to be predicted have been known, the data origins of these variables have been identified and the strategy to organize these data is defined, the next step of the procedure focuses on the generation of a data history. Automatic data collection strategies will be developed and as many samples as possible will be collected, discarding those with incomplete data.

5.3 Data Validation

Both the quality of the data and that the samples with which the model will be trained should be balanced in good / bad cases to avoid the model overfitting.

5.3.1 Data quality

An assessment of the data quality KPIs is carried out. These KPIs are:

- Completeness, which verifies that there is a value in all the fields.
- Compliance, which verifies that the data is in a standard, readable format.
- Consistency, which verifies that the data is not inconsistent.
- Precision, which verifies that the data takes the value it is supposed to take.
- Duplicity, which verifies that the data appears only once in the data set.
- Integrity, which verifies that all the desired data is available.

The objective of this profiling of the data is to find out if its state is adequate to train the model and, therefore, detect what needs to be corrected and, in turn, determine the control parameters that help measure progress in the tests.

5.3.2 Data balancing

To carry out a correct training of the model, the samples with good and bad quality must be balanced in the dataset. If the data set has not enough samples of good quality or of bad quality, techniques such as subsampling or oversampling can be applied, with which the balance between cases can be adjusted.

5.4 Development of prototype models

Tests are developed in a rapid model prototyping program, such as WEKA, using offline data and creating different combinations of use and organization of the data so as to see which combination and algorithm gives the best result. As humans, we like to use what we know. And we really like to use techniques we're good at. This can cause a problem in predictive modelling since different data types respond better to different modelling techniques. An application example is presented in Figure 5.2.

Metrics that are used to measure the model quality are:

- Confusion Matrix
- F1 Score
- Gain and Lift Charts
- Kolmogorov Smirnov Chart
- Area Under the ROC curve
- Log Loss
- Gini Coefficient
- Concordant – Discordant Ratio

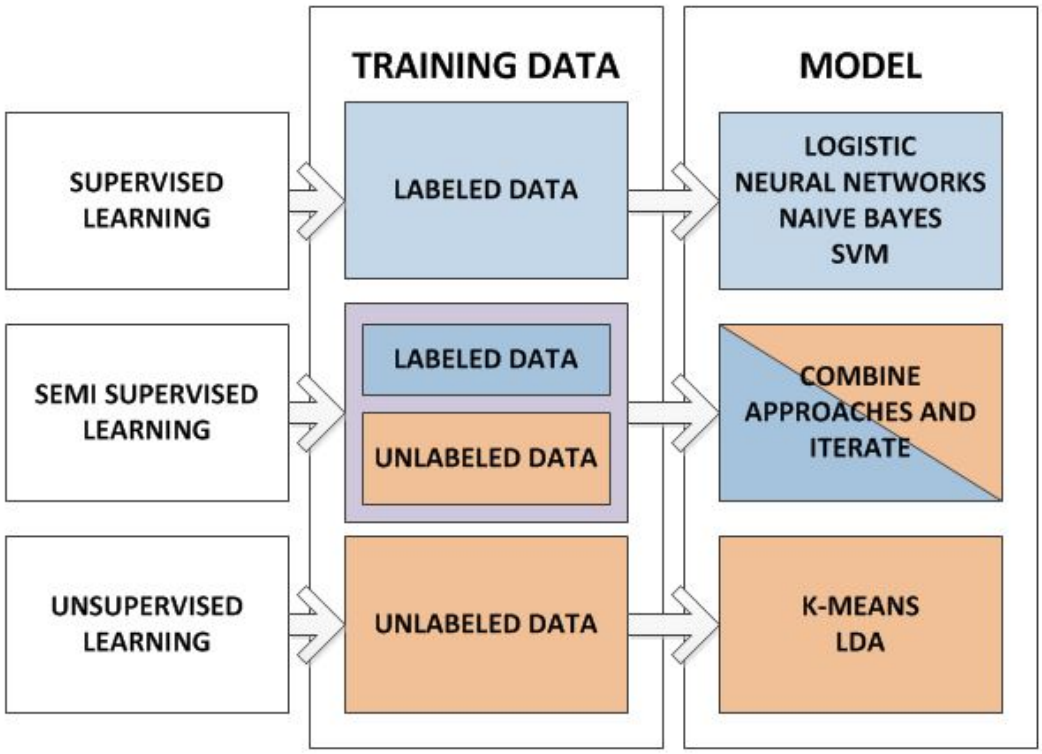


FIGURE 5.2: Prototype model development according to data type.

- Root Mean Squared Error (RMSE or RMSLE)
- R-Squared/Adjusted R-Squared

5.5 Development of the selected model

Once the optimal combination of data is known, both obtained from the process and generated through feature engineering, and prediction algorithms, the model is developed in a programming language that is compatible with plant systems, such as R, in such a way that can be exploited directly but maintaining the flexibility and control of actions of the programmed models. Model quality metrics of this new model are performed and compared with the previous case, with the same data set, it should perform as good as the model developed in WEKA.

5.6 Validation and deployment of the model

The solution is subjected to predefined acceptance criteria and, if accepted, will be integrated into the plant systems, from the online data extraction, transformation and loading systems, dataset generation systems, the model predictive and the selected output application, data visualization, messaging.

CHAPTER 6

Case study: Mercedes Benz Fábrica Vitoria

THE origins of the Mercedes Benz Fábrica Vitoria date back more than 60 years, being the oldest van production site in continental Europe. The factory, located on the outskirts of the city of Vitoria (Spain), occupies an area of about 600 000 square meters, with 257 000 of them destined for production. Likewise, it has 5 000 employees working three shifts and has a production capacity of more than 600 vans a day. These vans can be configured in a multitude of variants, from the powertrain (combustion or electric motor), body lengths (short, standard or long), propulsion types (front, rear or integral) or equipment levels, making practically no two vans alike. Among these variables that make up the van, the body colour is very remarkable. The van can be ordered in a wide variety of colours, from basic white to metallic special paint, varying the range of available colours every year and including the possibility of choosing unique decorations for fleets or special customers.

The factory is divided into four parts:

- **Body in white:** in this first part, the sheet metal forming is done to shape the body parts and then these parts are welded together composing the frame in which moving parts, powertrain, drivetrain, interior trim, seats and other equipment will be mounted.
- **Paint shop:** is the part of the process where all the necessary operations to apply a protective and coloured layer to the body in white. This part is of vital importance to guarantee the corrosion resistance of the body and ensure a high quality perceived by the customer. It is in this part where the research is focused.
- **Final assembly:** it is the final part of the van's manufacturing chain, where all the necessary components, mentioned before (powertrain, drivetrain and so on), are assembled.
- **Supplier park / I-Park:** it is the hall where the main suppliers of the factory are located, such as Grupo Antolín or Faurecia. It is located within the facilities of the plant and has and area of about 20 000 square meters.

In this sixth section, and industrial case is presented as a validation method of the previously presented procedures, following the steps of:

- Data set definition (Chapter 3)
- Data set generation (Chapter 4)
- Data set analysis (Chapter 5)

and adding specific paragraphs referred to

- Results
- Conclusions
- Improvement proposals

6.1 Experiment

The painting process corresponds to the standard described in the introduction section of this document. The considered paint shop workflow for this experiment is shown in Figure 6.1.

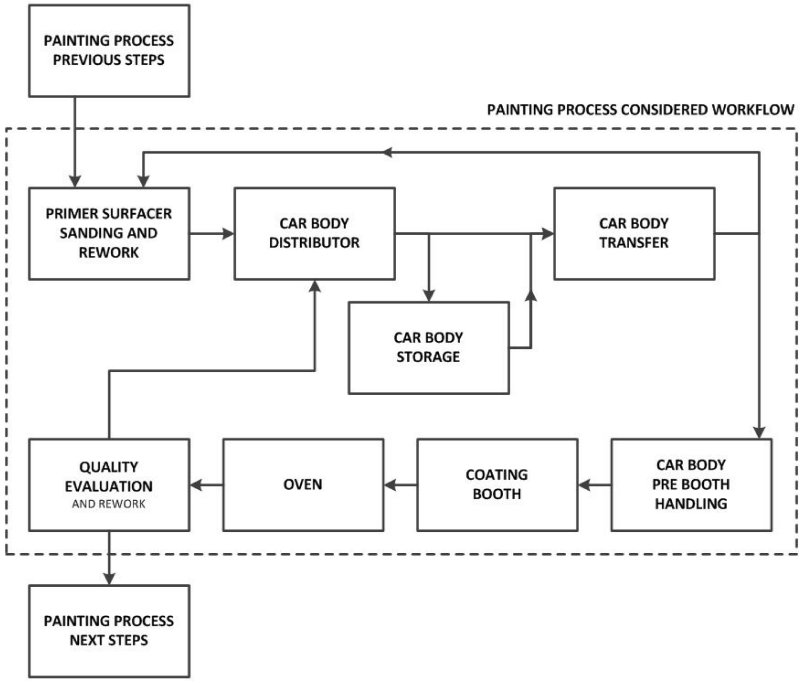


FIGURE 6.1: Paint shop workflow considered in the experiment.

Figure 6.1 shows the paint shop process activities that are going to be considered in order to predict quality outcome. The quality of the product up to this point in the process is assumed to be without defects. Although this supposes suppressing the previous variability, it is carried out as a simplification of the experiment since its motivation is to study the predictive capacity of the variables in this workflow.

The samples considered for the experiment are the metal parts that make up each vehicle body. This is because a distinction in the variants of the vehicle's chassis configuration that are possible is not considered, such as, the number of leaves of the tailgate (which has two options), the presence or absence of side sliding doors

or the installation of a panoramic roof. So, for example, if a chassis has a panoramic roof (the roof-metal-sheet has a cut all along the piece), the roof-piece of that chassis will not be part of the population of the group of roof-pieces. Another reason for this piece-level evaluation is that due to the rigorous quality controls motivated by the demand of *total quality* Chenhall, 1997 requirement, all the samples considered at the single-ID chassis level have defects (negative quality example), so it compromises viability when training the model. The automatic paint film quality evaluation divides the results of each chassis into 54 pieces (40 are external body surfaces, 14 internal). Only the 40 external results are used, as the internal surfaces evaluation is not completely automated. Thus, 40 models have been made and 40 prediction results have been obtained of which the most significant ones are shown. Another relevant consideration is that only vehicles painted in an specific silver enamel have been taken into account since it is, by far, the most applied color in this particular paint shop process and discards the variability due to the chemistry of each enamel.

Therefore, the variable regarding the color is not used. The number of process variables used is the maximum available on the dates of the experiment. There are three rounds of analysis with variations in the number of samples used to train the model, in the first round, 2697 chassis id are used, in the second round 27243 and in the third round 68558. Weka is used as the model developing software and multiple algorithms are considered. The best performance algorithms have been, decision tree in the first and second round and simple logistic regression in the third round.

The experiment consists of three stages, the data set definition, the data set generation and data analysis.

6.1.1 Data set definition

A series of steps are carried out in this first stage, such as: understanding the process, studying its production steps and the sequence of actions that make up the body painting process. Identify the *key performance indicators* (KPI) of the quality criteria that are the objective of the predictive model, as well as the process variables that influence them. For each step, the process input variables and disturbances that can influence the KPI of the quality expected in that sub process are identified. For the identification of the relevant variables for the prediction, a functional analysis of the process is performed, identifying the workflow, defined as the main steps, subprocesses, and actions that, without being a relevant process activity, may have an influence on the result to be evaluated.

A study of the factory documentation of the process, documentation about the control systems, a study about the state of the art of the process (Streitberger and Dossel, 2008) and the expert knowledge is necessary to develop the process analysis. As an example, for this experiment, the workflow analysis is carried out by standing at the beginning of the production line and following the car body throughout its painting process. Information is collected for each step of the process, identifying, through the knowledge sources defined above, the relevant variables, the disturbances and an evaluation of the process step performance related to, in this experiment, paint film quality parameters.

Another additional action to deploy in this first step is to develop a procedure for the univocal identification of variables. This is because the reference to each variable by any work team involved in the project must be clear. In this particular case, a

procedure for naming variables has been developed depending on the location of the measurement. This procedure is shown in Figure 6.2.

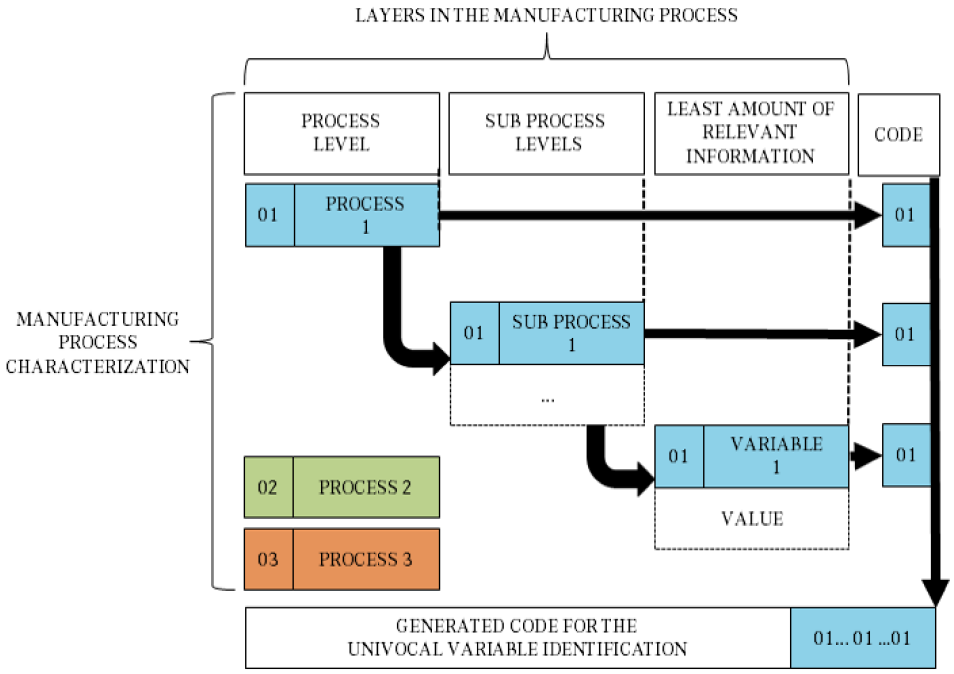


FIGURE 6.2: Variable unique ID generation procedure.

Figure 6.2 shows the code designation that is dependent on the location and has been arranged using the study of the process, by identifying the main processes that form the painting process, and within each as layers, the sub processes, functional groups and equipment until reaching the measure to be identified. The classification has been done numerically, also assigning colors with different degradations to each process and related sub processes and machinery.

6.1.2 Data set generation

The objective of the data generation procedure is to build a data board in which the rows will be each unique chassis and each column is one of the variables considered relevant to the process (see Figure 6.3).

Each chassis number represents a group of process data with the values of the variables at the time they affected the corresponding bodywork. In this study, it has not been stored in relation to the pair chassis-piece since it has not been able to perform such a precise synchronization, with what the values assigned to each body, have been to each of its parts. Both categorical and numeric variables are recorded. The numerical values stored in each box of the data board correspond to the entire time series defined for each variable. In this experiment, the measure of the statistical mean is used to represent each group of values.

To obtain the values of the variables that were identified in the data set definition procedure, a search of the data sources is carried out. Here, the identified variables that are considered relevant for the model by the expert knowledge are filtered in order to save efforts in the subsequent procedural steps. This experiment considers a source of data to any form of data acquisition and registration of the identified variables. Taking into account the characteristics of each data source, a classification

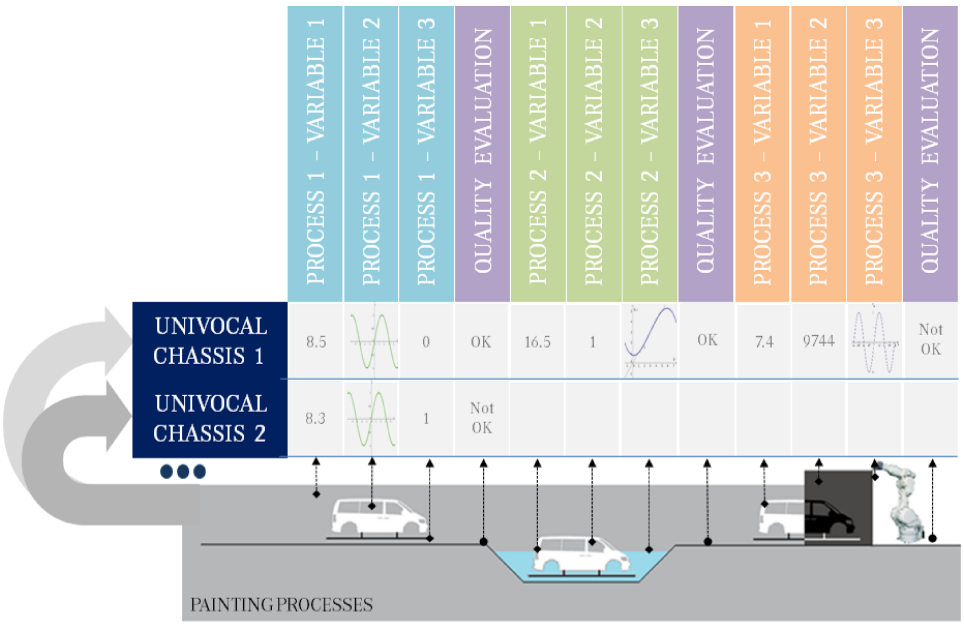


FIGURE 6.3: Data set synchronisation procedure.

of the relevant variables is made according to several measurement status considerations (see Figure 6.4).

Firstly, a difference is made between measured variables, by any means, or not measured. As a consequence of this first filtering, procedures that lead to obtaining data from the not measured variables are developed, as, for example, the study of the physical or logical dimension of the variable in order to define the appropriate measuring method, or, the consideration of proxy variables that are already available. Secondly, from the measured, a differentiation is made attending to the variable measurement digitization, if the variable sensor or indicator is mechanic or digital, so the digital data is somewhere within the systems. This lead to define procedures for signal digitization. In the third and last filter, digital measurement availability is considered. This evaluates whether the data can be downloaded or is in a unreachable environment. Thus, methodologies for extracting data from such unreachable sources can be studied (Wang and Qi, 2009-Zhu et al., 2015).

Once the already available data sources are identified, their data are synchronized. The primary synchronization key is the chassis number (Grundig and Klein, 2016) and the synchronization criteria, as mentioned above, is to store the values of the variables while they affect the bodywork. In the painting plant of the experiment, data capture and storage systems contain different information fields. Because of this, the synchronization has been done by assembling the available data in each usable data source. To achieve this synchronization sequence, a special algorithm has been developed that consults the production data to know the painted bodies and their color characteristics and their production times, referred to the times of passage through control points and stay times in stations. Thus, the timestamps of the variables stored in other data sources are identified so they are assigned to the chassis at the corresponding time. This information is completed with the paint film quality data, so that quality results are linked to their corresponding process data (see Figure 6.5).

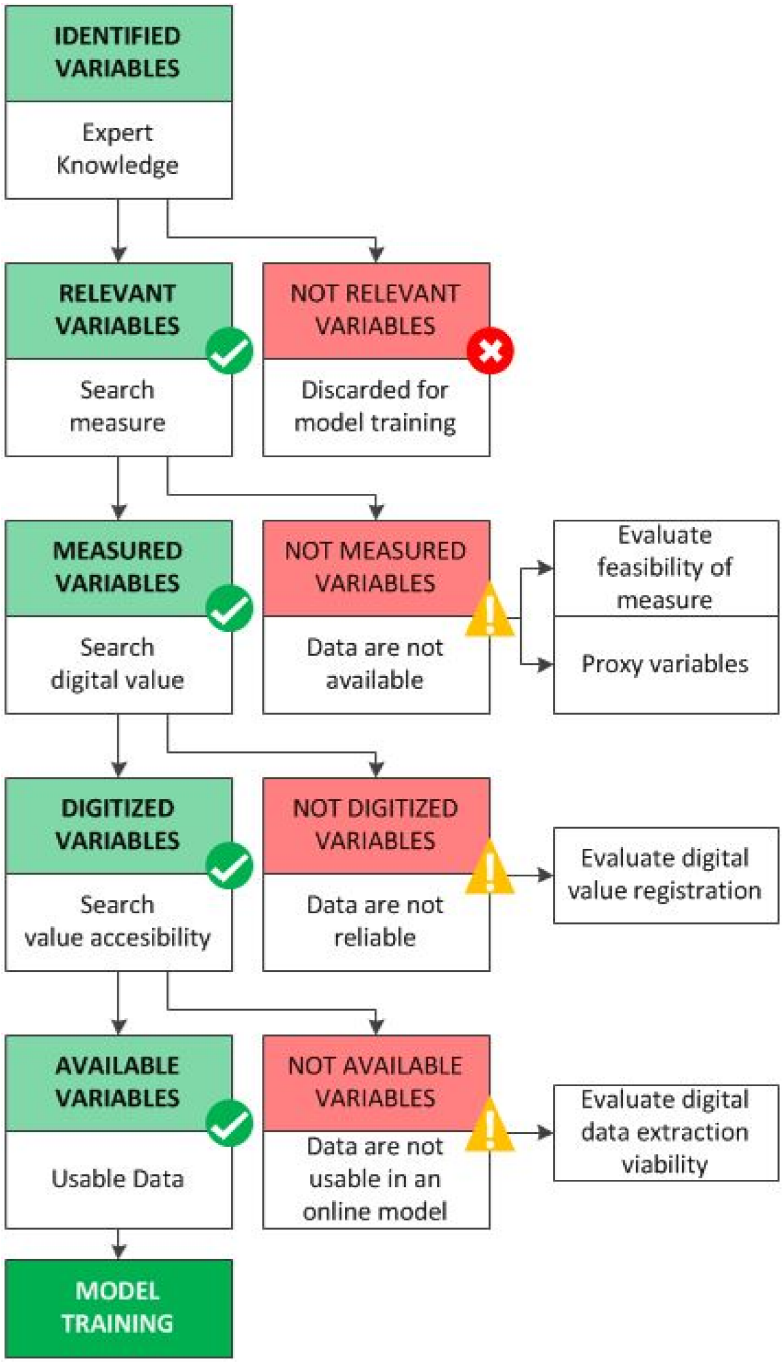


FIGURE 6.4: Variable classification.

6.1.3 Data analysis

Once the data of the variables are synchronized with the chassis ID as the primary key, a series of steps are carried out prior to the training of the models. In this experiment, an evaluation of the quality of the sample's data has been made (Press, 2016) and the balancing of the samples' good/bad quality cases has been evaluated.

First, to validate the data of the samples, a series of verified indicators have been taken into account for the data quality of the different data sources Pipino, Lee, and Wang, 2002. The *completeness* consists of counting the number of null values in the data source, indicated as a percentage of the total. The *conformance* is established by

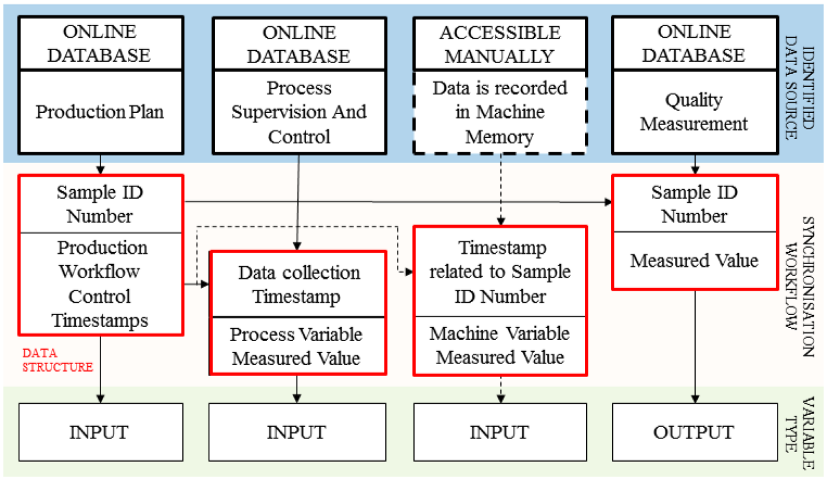


FIGURE 6.5: Use case data set synchronisation procedure.

two steps, first, establish the correct format for each variable for, second, then post the number of instances in the correct format. The *consistency* is to check if the value studied is far from the average of the normal values. The *accuracy/precision* studies the dispersion of the sensor values. The *duplicity* takes into account the number of duplicate data among the files that are data sources. The *integrity* determines for each line the expected variables, and it is verified that we have all the expected ones. A car body whose data variables do not meet the data quality criteria has been discarded. This led the experiment to have a very relevant loss of samples at first (see Figure 6.6) since incomplete or non-conforming data have appeared as the enamel coating process progressed (the enamel coating process of the paint shop analysed in the experiment is divided into BC1-BC2-BC3-CC1-CC2 named sub-processes). Once the first data quality problem appears, the sample was already discarded. Subsequently, data quality improvement procedures were implemented which managed to correct the situation to a great extent (see Figure 6.7).

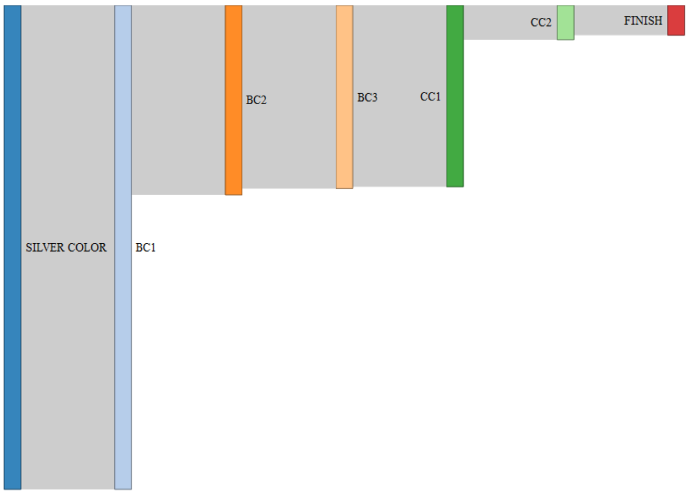


FIGURE 6.6: Valid samples before data quality enhancement.

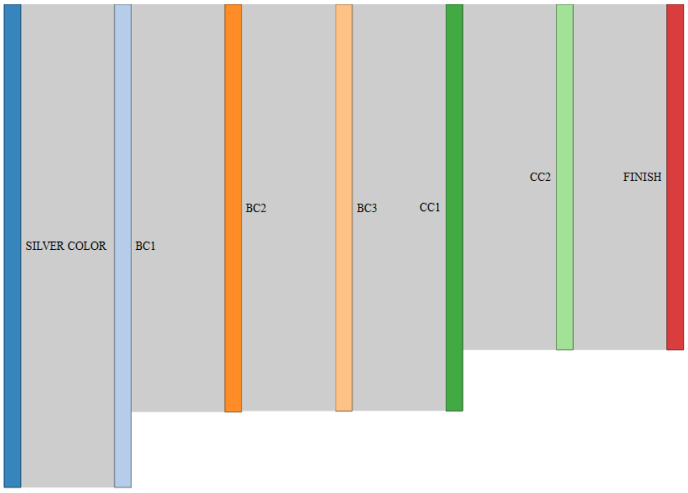


FIGURE 6.7: Valid samples after data quality enhancement.

As mentioned in previous sections, each sample, which in the table corresponds to a unique chassis ID, has been decomposed in 40 parts. A second step prior to the training of the model has been the evaluation of the balancing of positive/negative quality evaluation in the samples (corresponding to each said part ID). Oversampling techniques as SMOTE (Chawla et al., 2002) have been applied to balance the data sets to the maximum. Good data set balancing results have been obtained in some of the part ID's (green marked cases in Figure 6.8). The model training efforts are applied to these parts (see Table 6.1).

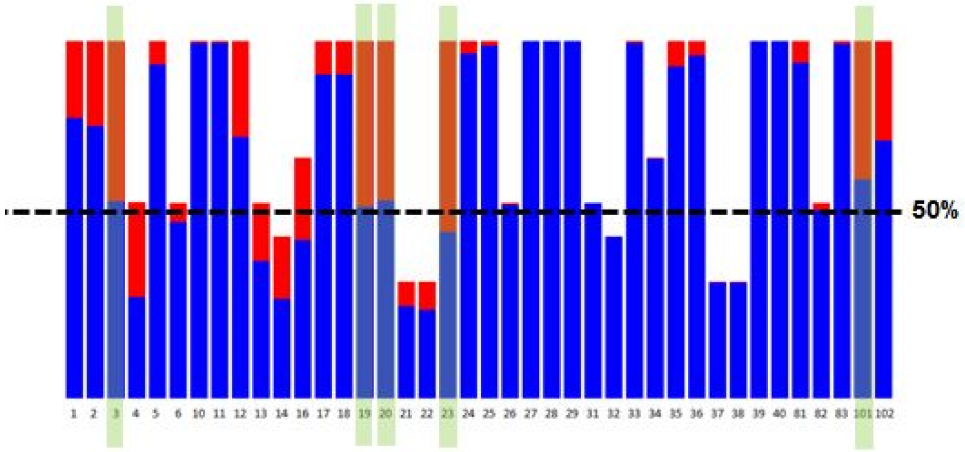


FIGURE 6.8: Sample good/bad quality distribution. Per PartID

TABLE 6.1: Vehicle parts codification.

PART ID	PART DEFINITION
3	External-right side sliding door hinge.
19	Front right door.
20	Front left door.
23	External side hood.
101	Front right door. Internal side cavity w/out post

Once the data have been separated and validated in the corresponding data sets, a series of Supervised Machine Learning Algorithms are developed (Wolpert, 1996-Wolpert, Macready, et al., 1997). In this experiment, the algorithms were developed with the *WEKA* (Garner et al., 1995-Hall et al., 2009) software as it allows rapid model prototyping. Precision Steyerberg et al., 2003 and area under ROC curve (AUROC)Fawcett, 2006 metrics are used to evaluate the model performance.

6.2 Results

The results of the experiment are presented at several levels: at the level of the definition, it has contributed to the improvement of the knowledge about the process. An analysis of the process has been made, providing, for each step of the same, the relevant variables, disturbances and expected results of each sub process. A variable definition limit has also been identified, in which the denomination of each variable by each plant team is made according to their particular interests (see Figure 6.9). To overcome the limit, a unique coding system has been proposed for the identified variables (see Figure 6.2).

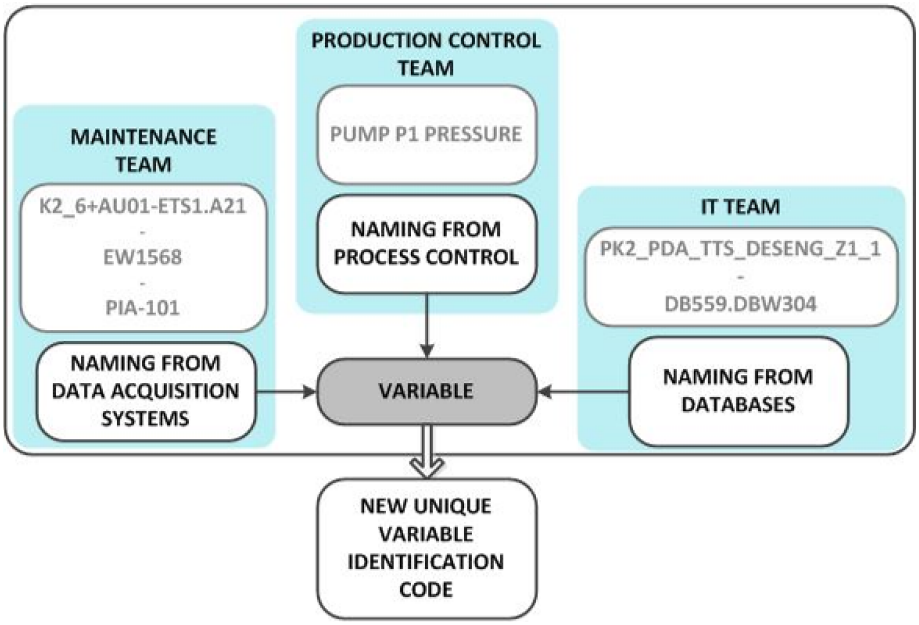


FIGURE 6.9: Variable unique identification.

In the generation, the data sources of the plant have been analysed, searching the data of the variables and making a classification of them according to the availability of their data (see Table 6.2). In this way, procedures have been developed so that all

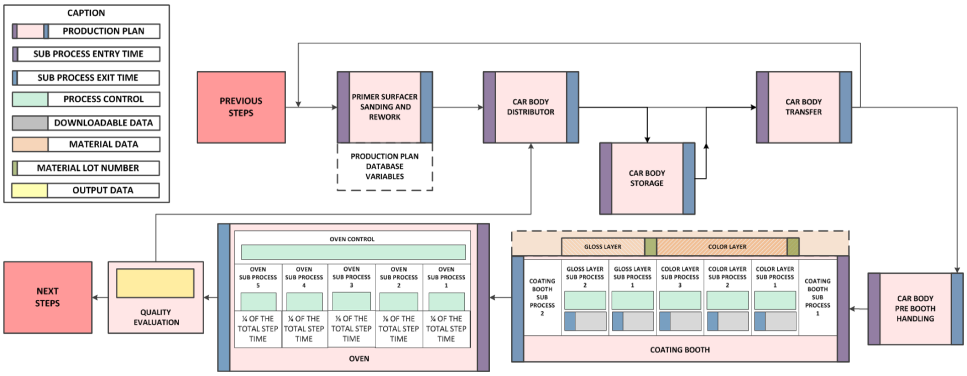


FIGURE 6.10: Synchronization procedure of experiment data sources.

data, whatever their origin, will be available in the near future. These generation procedures have been based on a study of the state of the art in the generation of data from the painting process within the Daimler group.
For synchronization, a schema has been generated (see Figure 6.10) that orders, in the process sequence (spatial synchronization), all the digitized and available data sources, so that the data can be assigned to the variables according to the primary key, the chassis number, as accurately as possible (temporary synchronization).

TABLE 6.2: Experiment variables.

	AMOUNT	PERCENTAGE
Identified Variables	679	100
Relevant Variables	402	59
Measured Variables	349	51
Digitized Variables	121	18
Available Variables	104	15

Data have been obtained for 15% of the variables identified (see Table 6.2) as relevant. In these 104 variables used to train the model, environmental variables (temperatures and humidity), production variables (such as dwell times of the bodies in the sub processes, the passage through unfinished body storage areas or oven time) or application variables (such as paint product volumes applied by the robots) are included.

As results of the analysis, precision and AUROC have been obtained for three rounds of data. The difference is the number of samples used to train the model. Related to the number of samples, another valuable result is the development of procedures that improve data quality and allow to have as many useful samples as possible. The model results have been considered for the part ID with the best balanced datasets,

which are parts 3, 19, 20, 23 and 101. In the first and second round, the best performing algorithm has been the decision tree and in the third one a simple logistic regression. The most promising results have been in part 20, with a precision of 65% and an AUROC of 0.69 (see Table 6.3).

TABLE 6.3: Model results.

	FIRST ROUND *2697 samples*		**SECOND ROUND** *27243 samples*		**THIRD ROUND** *68558 samples*	
PART ID	PRECISION	AUROC	PRECISION	AUROC	PRECISION	AUROC
3	50.58%	0.511	60.52%	0.602	62.72%	0.681
19	50.92%	0.518	55.89%	0.550	61.40%	0.671
20	52.18%	0.520	61.18%	0.652	65.03%	0.690
23	49.69%	0.507	53.56%	0.537	58.48%	0.604
101	65.86%	0.521	71.19%	0.554	66.07%	0.672

The levels of accuracy and AUROC have been increasing as the number of samples increased (see Figure 6.11) except for the piece 101, in which accuracy has dropped between the second and third round of analysis despite increasing the AUROC.

FIGURE 6.11: Graphic representation of results.

In the third round, the variables considered most relevant for the prediction of the model have been, in general, the application variables (see Figure 6.12) have been considered as the ones with the highest predictive potential among the data set.

This makes sense since, for example, an excess in the amount of material applied could increase the risk of contamination due to the atomized particles or, a lack of material could could worsen the surface coating capabilities of the applied paint layer, making surface defects easier to see. As an additional detail to check the validity of the results of the model, looking at the results of the part number 20 (see Figure 6.12), the left front door, the variables of applied material correspond to robots of application of that same side, the left side, which can be considered as coherent.

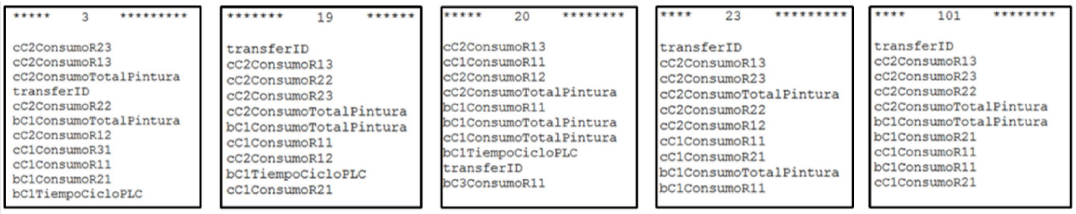


FIGURE 6.12: Variables with the highest predictive potential.

6.3 Discussion

The quality issues considered in this experiment are dirt and crater. Dirt can appear because of common causes and, mainly, because of special causes Taleb, 2007. So, as a defect to be predicted by the model, it has high uncertainty. The reason why it was chosen is because it is the only paint film quality issue whose apparition is common (over 90% of paint film issues are caused by inclusions as dirt) and could contribute to generate a significant data set to train the model. As a comparison, the second appearing issue are craters and are about a 3%.

Another reason that gives uncertainty to the model is that the defects that have been inherited from previous stages of the painting process are not considered. The automatic evaluation of defects is not exhaustive enough to determine in which layer of the paint the contamination, or the origin of the problem, has occurred. This is due to the fact that reaching this point is a highly resource-intensive activity that usually involves carrying out destructive tests. Improving the classification of defects by discriminating the layer in which they have occurred is, clearly, one aspect in which to carry out more research and that will contribute to improving the decision making of a predictive quality model applied to the entire car body painting process. That is why, because it is the last layer, it has been decided to carry out the experiment in the enamel process, assuming that all defects have appeared in the colour and bright layer, without inheriting anything from previous stages.

Considering the scarce percentage of relevant variables that have been used for training, the model behaves reasonably well, making it clear that there is margin for improvement if the number of used variables increases.

The experiment has also been useful to evaluate the digitization degree of the paint shop Bley, Leyh, and Schäffer, 2016. The limits in the information management systems of the plant have been detected and have enable to elaborate procedures in order to take steps towards the development of a data-driven decision-making system.

6.4 Conclusions

As a conclusion about the results, it can be inferred that there is a correlation between the input data of the process and the paint film quality output so the objective of implementing a predictive control in the plant is feasible. Regarding the model, it is able to identify variables that are relevant for the KPI. These variables are related to the amount of paint used by the application robots in the paint booth mainly because:

- An excessive amount of paint can cause problems of sags or over-spray contamination.

- An amount of paint smaller than necessary causes a reduction in coating capability.

It should be pointed out that in both cases these values are within the operating regime of the machine, otherwise, values out of the defined range cause an error. As these errors are critical, those chassis IDs with these said errors are not used to train the model.

In addition, it is verified that there is a relation between the amount of paint that is applied in the coating booth and the paint film quality because the body pieces on the left are more affected by the paint applicators on the left and vice versa.

Increasing the sample size improves both the accuracy and the AUROC of the model. It must be taken into account that the relation between the increase in the sample size and the improvement in accuracy is not linear, since the increase in accuracy caused by the addition of a new sample decreases as the sample size is larger. There is also a limit on the predictive capacity of the model as increasing the accuracy requires an exaggerated number of samples. This implies that it would only be possible to significantly increase the accuracy of the model by including additional input variables, selecting them from those identified but not used yet.

As mentioned above, not all the identified input variables were accessible at the time of the research and, therefore, it was not possible to collect their values, either because there was no measurement system or the data collection system was not appropriate. It is therefore concluded that in order to improve the accuracy of the model it is essential to increase the degree of digitalization of the plant so that values from all, or at least a greater percentage, of the relevant identified variables can be extracted. It should be taken into account that, even if values were obtained for all the relevant variables, the painting process presents a great variability since it is subjected to what is known as special causes, i. e., uncontrollable events, such as, contamination due to changes in the atmosphere of the paint booth.

It should be remembered that, in this research, the defects inherited from the previous processes to the enamel coating have not been considered, since in these previous steps, the quality controls are not as exhaustive as the final quality control and many of these defects can only be discovered through destructive tests that are performed only if a defect appears on a recurring basis. The limitation introduced by this simplification could only be avoided by measuring the quality of each of the previous layers with the same level of detail, both in terms of the number of samples and the number of registered variables.

As a resume, from the definition phase, it is concluded that:

- In a complex environment such as an industrial plant, with diverse nature data origins, it is necessary to develop a methodology of univocal denomination of variables so that any worker of the plant is able to easily locate the data source that produces each of the variables, regardless of its nature.

From the generation phase it is concluded that in order to obtain sufficient data to successfully train a predictive model of paint film quality, it is necessary that:

- The factory data is accessible, in terms of the number of variables measured and the speed of data collection.

- The spacial and temporal synchronization of the variables is possible, that is, the possibility of identifying the location where a variable occurs and the capability of selecting the precise timestamp of the variable when its value was relevant for the quality of the paint layer.

Finally, from the analysis phase, it is concluded that:

- It is mandatory to guarantee the quality of the data. Thus, it is necessary to define a series of parameters (completeness, conformance, consistency, accuracy, duplicity and integrity) that data must meet to a defined degree so as to be accepted as valid data for training the model.
- The number of training samples must be balanced. In this case study, with regard to cases of good and poor quality, so that the model is not skewed towards the case overrepresented in the sample.

Analysing the results of precision and AUROC values obtained from the prediction model of quality considering different sample sizes (in the first round, 2 697 valid samples are used, in the second round 27 243 and in the third round 68 558) and supervised machine learning algorithms (the best performing algorithms were decision tree in the first and second round and simple logistic regression in the third round), it is concluded that the correlation between the input and output data exists. Therefore, it is possible to implement a predictive control in the paint shop that helps improve the efficiency of the process, in terms of rework savings, while maintaining the product quality. It should be mentioned that to avoid the limitations found through the research and improve the results of precision and AUROC of the predictive model, it is necessary to increase the number of available variable values on the generation of industrial data on legacy machinery, not designed for data streaming in first place. Also, further work is needed with regard to the spatial and temporal synchronization of the process data that is generated from different types of sources, so that the precision with which the measurements are assigned to the variables in the data set is increased.

6.5 Proposals in order to improve the results of the predictive model in Fábrica Vitoria

Once the case study results are presented and predictive model improvement opportunities are identified, applying the knowledge gathered from the analysis of the state of the art in data generation, this section aims to present a series of proposals for the generation and synchronization of process data for the Mercedes Benz Fábrica Vitoria paint shop.

Regarding the generation of process data, the objectives are:

- Once the data generation structure of the paint shop is defined, propose appropriate data generation methods for the process data acquisition.
- Propose the data generation architecture that supports the generation of process data. The generation architecture that is proposed can be integrated into the data management systems of Fábrica Vitoria

Regarding the synchronization of process data, the objective is to propose a method for the synchronization of variables. This method is based on Timestamps and allows a synchronization of process variables in which each car body is going to be

the primary synchronization key of the variables.

6.5.1 Data extraction proposals

As mentioned before, data sources were characterized as non digitized, digitized but not usable and usable. Attending the not digitized or not usable variables some actions are proposed:

- Variables that are not measured or not measurable: it is proposed to analyse the cost of the digitized capture of the value of the variable or the use of proxy variables, variables whose values are available and from which the value of the variable of interest can be extrapolated. The evolution in the digitization of this case leads to the implantation of a digitized sensor for each variable. Figure 6.13 shows the action flow proposed for this case.
- Variables that are measured with a mechanic instrument, not digitized: analogously to the previous case, the incorporation of digitized sensors is proposed to be able to have the data of these variables. Figure 6.14 shows the action flow proposed for this case.
- Record sheets (paper sheet filled by hand): it is proposed that the data must be uploaded directly into the database through an user application or through a digital file that is then uploaded to the database. It must be taken into account that these solutions are used to obtain the information offline (referring to the fact that it is not possible to have the data in Real Time or Near Real Time, being preferable the generalized use of digitized sensors that make the data immediately available and without human mediation in order to obtain the variable's value with reliability.

6.5.2 Data generation architecture proposals

This section also proposes the data architecture for the process data generation system for the Mercedes Benz Fábrica Vitoria paint shop. This architecture is applicable to digitized and usable data. A summary of the proposed architecture is presented in Figure 6.15. It shows, from left to right, the parts of the generation system, which are: data ingestion, storage, processing and visualization.

Ingestion is referred to the data acquisition from the data sources and sending these data to storage and processing. The storage part has a real-time part and a batch processing part. The processing part can have an advanced control part, which can return results to the process. The visualization part is the one that represents decision making through data. Next, each step presented in Figure 6.15 is explained.

Ingestion

Ingestion includes the data sources of the process, the data generation methods and the ETL processes (extraction, transformation and loading), as shown in Figure 6.16.

Here, referring to the process, it is necessary to distinguish between data origin and generation point. The origin is the machine, the generation point is the place on the machine where the data is located (communication bus, memory...). Each generation method executes:

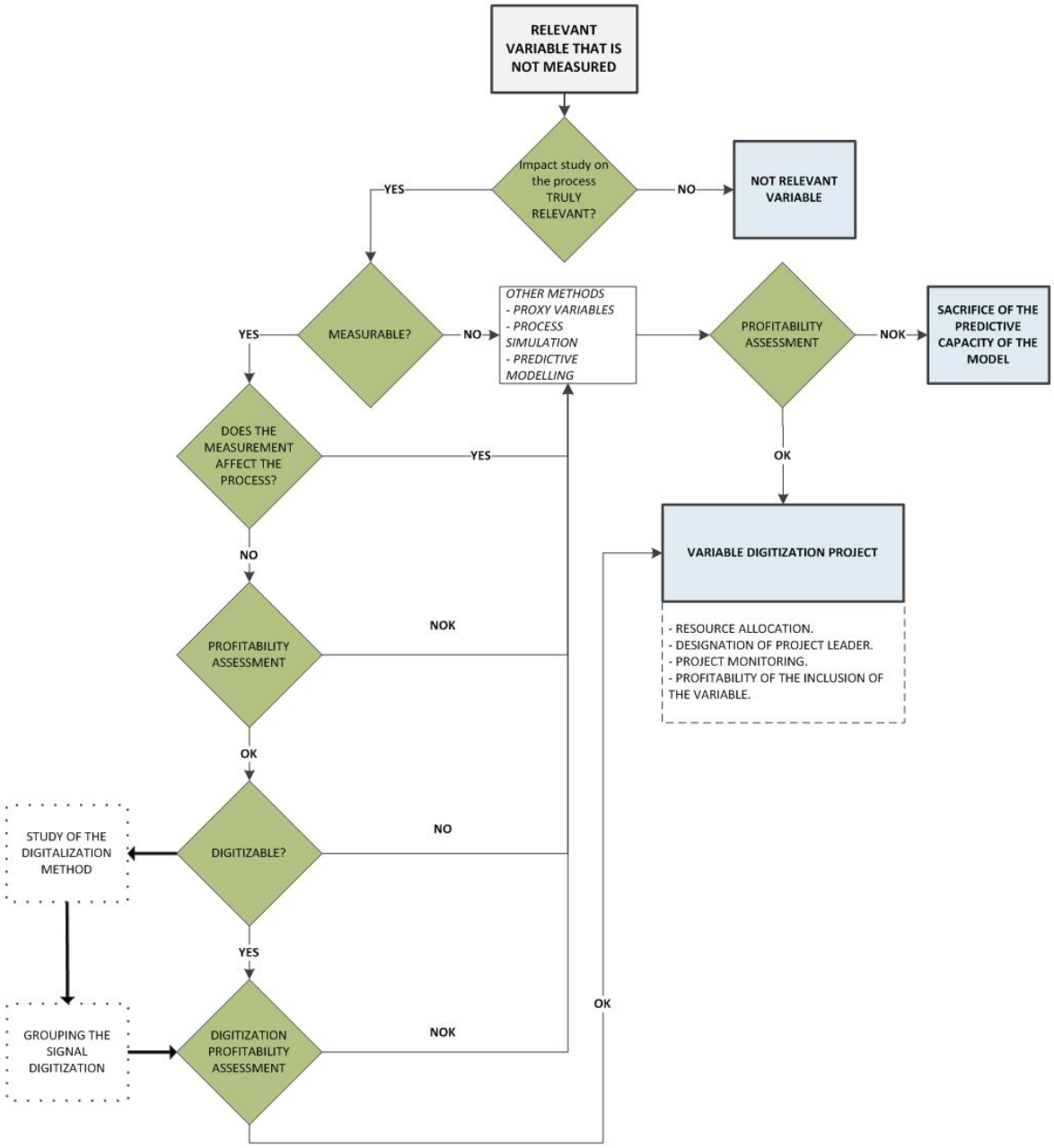


FIGURE 6.13: Action flow for identified relevant variables that are not measured.

- The acquisition of raw data (usable data) from the generation point corresponding to each data source.
- The translation of this exploitable data into usable data through the necessary code transformations (frames).
- Making this data available to the following systems of the architecture.

At this point, it should be noted that these data obtained from the machine can be used to carry out a primary control of the process through the implementation of simple rules for the generation of alarms, messages or visualization. This technique is known as Edge Computing. These acquired process data can be classified into

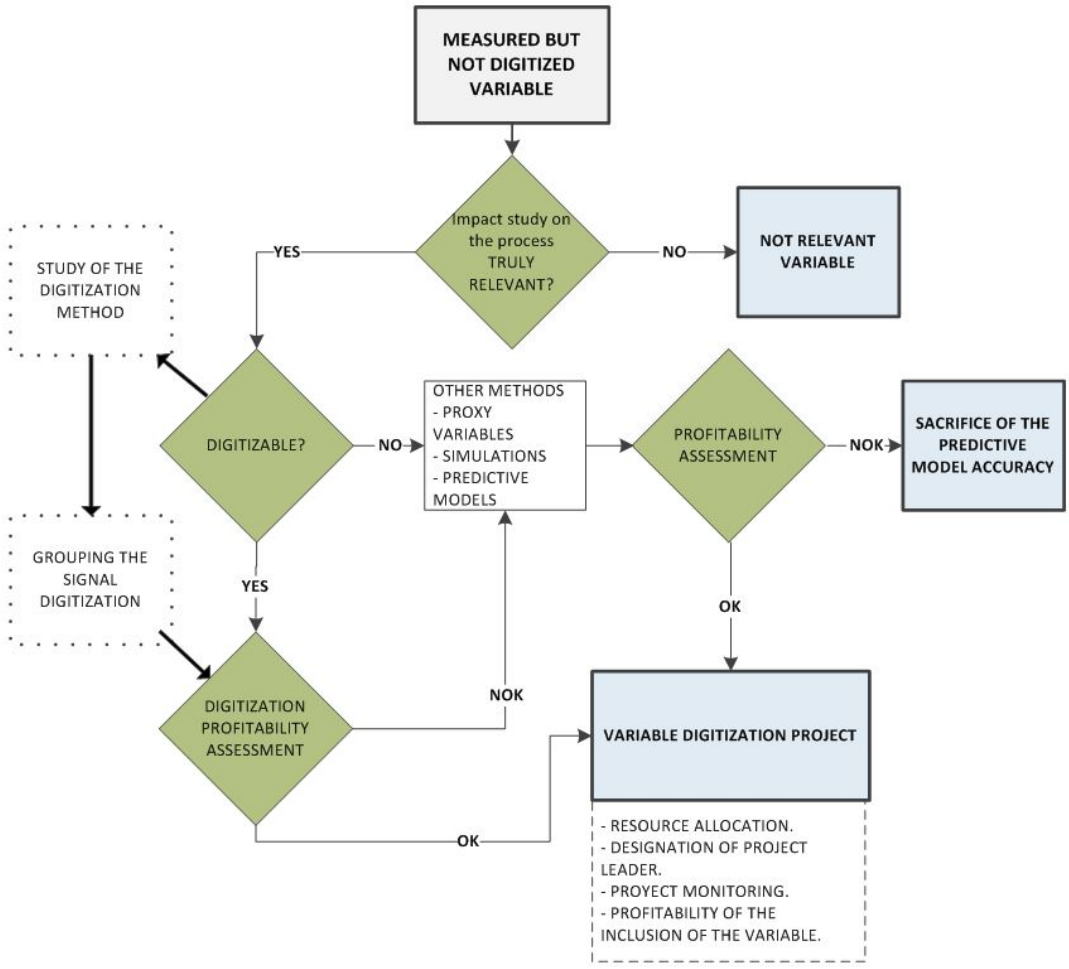


FIGURE 6.14: Action flow for measured variables that are not digitized.

several categories. For example, they can be classified according to the acquisition frequency, according to the generation point or according to its corresponding generation method. Thus, in the specific case of a PLC, if the generation point is the communications bus, data can be extracted at higher frequencies than those of the PLC scan cycle. However, if the point is the memory of the PLC itself (the PLC updates the values present in the analogic or digital inputs at the beginning of each scan cycle) the data sent over the bus for the rest of the cycle is lost. In this case, data acquisition through reading the PLC memory blocks or through an OPC interface, depends on the operating cycle of the PLC. Therefore, the acquisition frequency will be, at most, the same as that of the PLC scan cycle. Thus, data can be extracted from the same source but from different data extraction points, each with its corresponding method.

As a conclusion, in this example, referring to the data of a PLC, the data can be extracted "in real time" from the communication bus between PLC and sensors or actuators and the data in frequency equal to or greater than the PLC scan cycle from the memory blocks or the OPC interface, if available in the PLC. The next step in Figure 6.15 is the parts of ETL, storage, and processing.

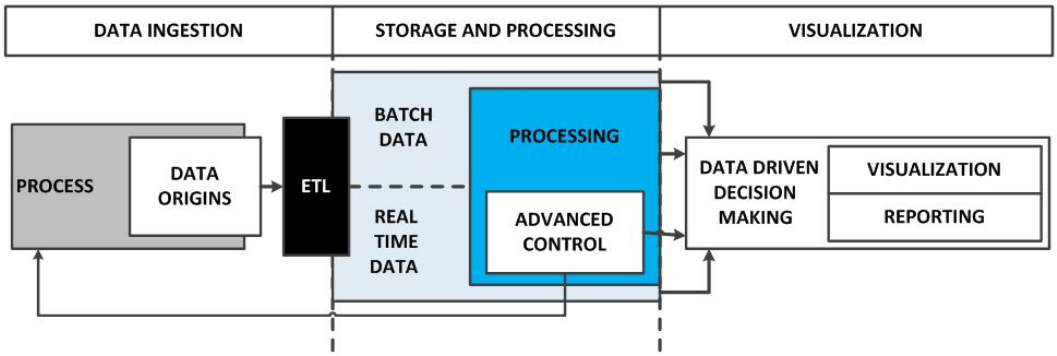


FIGURE 6.15: Generation architecture.

ETL, storage and processing

The part of the architecture dealing with storage and advanced processing is presented in Figure 6.17.

To centralize the collection, management and distribution of data, the use of an ETL system is proposed. The ETL system is made up of a set of software (in this case the NIFI and KAFKA software) to manage the flow of data generated from the different sources (data buses, PLC) to the next point of information consumption as is presented in Figure 6.15.

As can be seen in Figure 6.15, without an ETL system that centralizes data management between data sources and consumers, management must be carried out individually for each connection, specifying what data is needed by each consumer, the synchronization is done specifically for each connection and so on. With the ETL system, as shown in Figure 6.19, the individual connection assignment problem is solved.

As indicated in Figure 6.19, each data source sends the data to the broker (in this case it is the KAFKA software) and the broker makes it available to consumers through topics, to which these consumers must subscribe to obtain the data. In each topic, variables from different sources can coincide and the consumer can be a producer and insert the result back into KAFKA to be consumed by others. As ETL methods, the following are proposed:

- NIFI as a solution to guarantee the flow of data from the different sources, assigning the pertinent priorities and making the necessary adaptations to the data. NIFI supports messages of arbitrary sizes (message size on the order of GB or greater). This solution is integrable in SAP and has the following connectors for the identified data sources:
  - For SQL databases, the Query Database Table connector.
  - For non-SQL databases, NoSQL DB Processor or GetMongo.
  - For data that circulates on the TCP, UDP or Syslog network, the Network Listening Processor connector. Which corresponds to the PRIMAS, INTEGRA, SNAP7 and SENSORIC CAPTURE INFRASTRUCTURE methods.
  - For OPC, the OPC UA Processor connector.

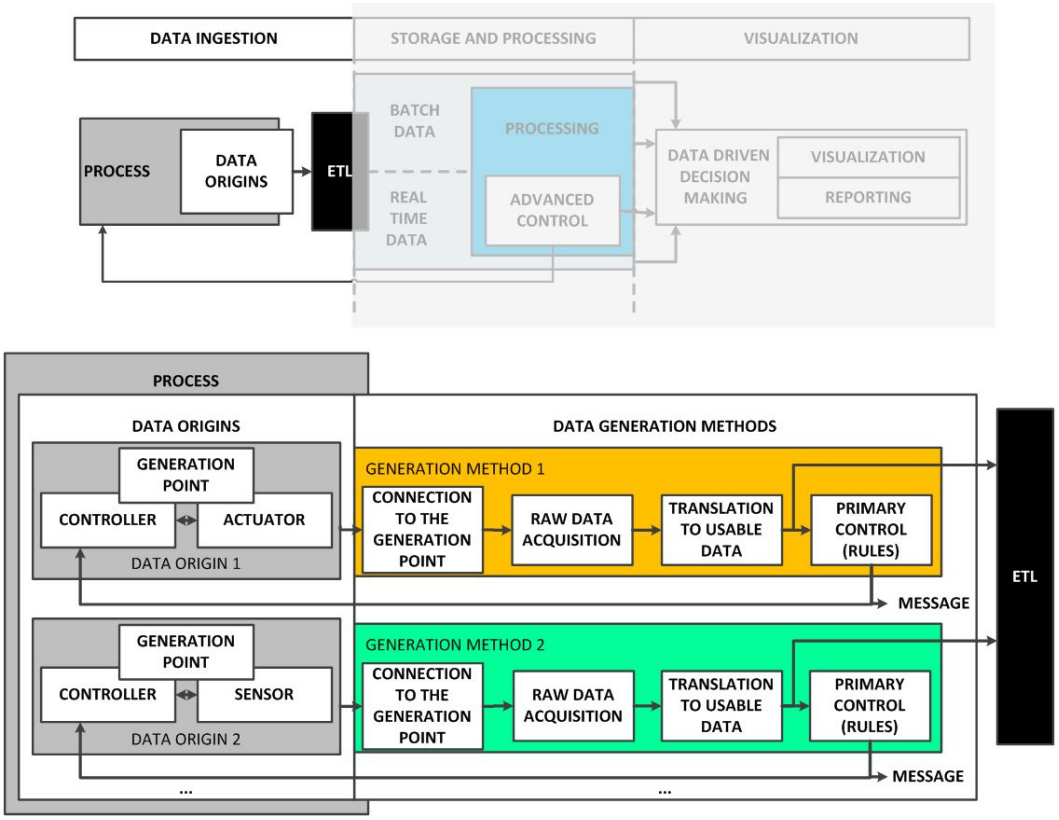


FIGURE 6.16: Ingestion in the proposed data architecture.

As a precaution in this part of data ingestion, it is important to consider the importance of maintaining the data structure of the data source, that is, that the order, size and characteristics that are considered in the ingestion, remain invariant, to limit negative variations in the quality measures of the data acquired.

- KAFKA as a solution to get the messages to the possible multiple consumers with a very low latency. The messages are in the order of KB to MB in size. KAFKA is integrable in SAP.

As can be seen in Figure 6.20, after the ETL part of the architecture, the batch and real time parts are presented.

In Figure 6.20, it is appreciated how the data travels from the data flow managers, ETL, and is offered to the applications (consumers) through a broker, which is in charge of managing the availability of this data. In this case, two consuming parts are distinguished, the batch part and the real time part.

- Batch is in charge of long-term storage where data is stored in a Data Hub. Unlike in a Data Lake, the data is labeled and ordered in order to avoid harmonization rework every time it is needed. Also, the possibilities of creating a Data Swamp are limited. The processing carried out in the batch part is oriented to routine tasks such as the generation of reports (in which a large amount of information will be used) or in activities with a high processing load but less frequent (such as re-training of the models used in the real time part). Here, the Hortonworks HDP data platform is distinguished, which is

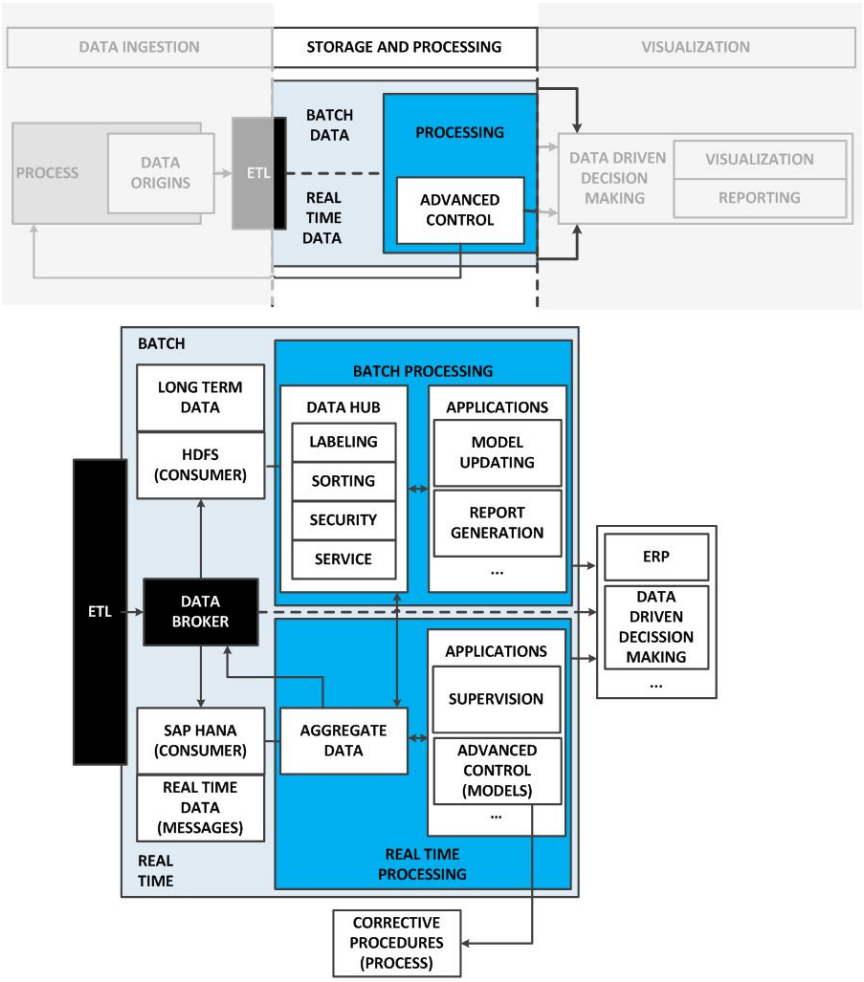


FIGURE 6.17: Storage and processing in the proposed data architecture.

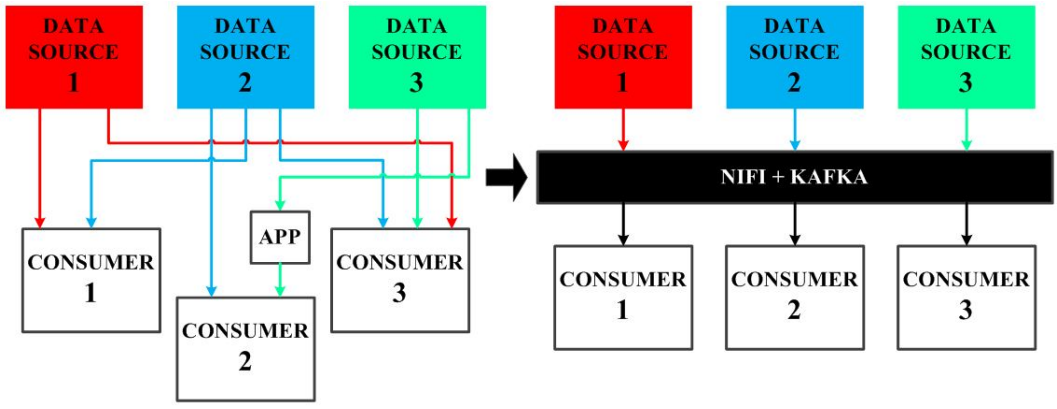


FIGURE 6.18: ETL in the proposed data architecture.

also compatible with SAP. In this platform, the Hadoop Distributed File System (HDFS) type storage is used and where the process data warehouse, Data Hub, is located. The processing part (YARN) performs tasks that are executed periodically and that uses large amounts of data, such as, generating reports

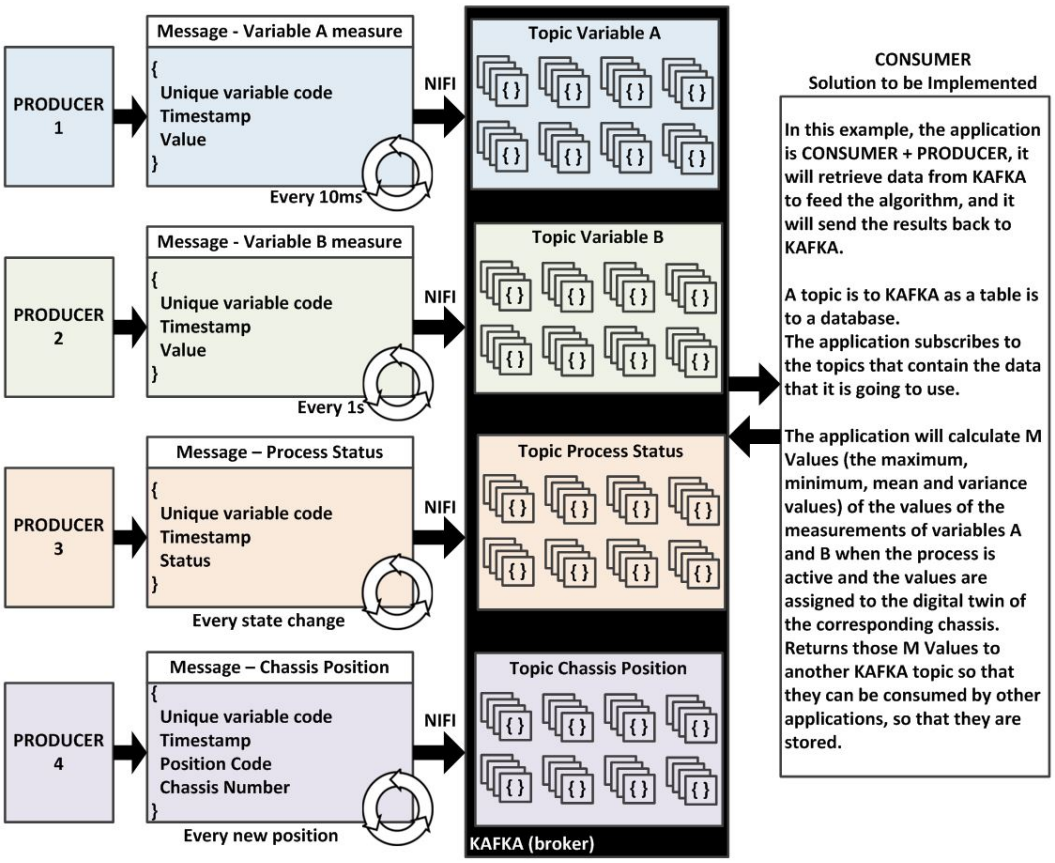


FIGURE 6.19: Producer-Broker-Consumer data architecture.

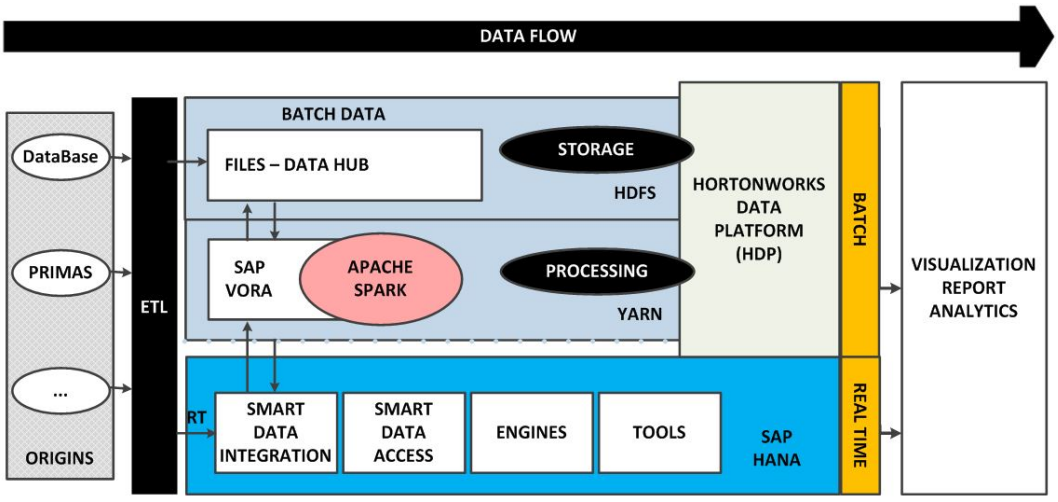


FIGURE 6.20: Data flow in the data architecture.

and training predictive models. For this, SAP VORA is used,it is a tool based on Apache SPARK, which is used for in-memory query optimization and data aggregation from various sources. SAP VORA allows to consume HDFS data and data from SAP HANA simultaneously.

- In the real-time part, the storage and processing parts are mixed. The data is

made available to applications in a data stream. These applications can have a multitude of functionalities and also, they can dump data generated by themselves to the data flow. Examples of applications can be: aggregated data generation applications, advanced control applications, such as predictive models, or monitoring applications (such as the product and process digital twin). SAP HANA software is located in this real-time part. Here the Smart Data Integration is different, which allows the integration of data in real time and in batch (from VORA). Smart Data Access allows access to remote data that is not copied to HANA as well as other tools and application engines. Subsequently, the SAP Cloud Platform tool is also distinguished. This is a HANA tool that is used as a form of data entry from applications developed in the cloud to, for example, replace the sheets of paper currently used to record process data and thus have the data directly in real-time.

Following the data flow presented in Figure 6.20, the next step is visualization. For the visualization part, a multitude of different softwares are available. In this architecture proposal the SAP ERP, the visualization system SAP Lumira or applications developed using SAP Fiori are used.

Data visualization

In the visualization part, Figure 6.21, applications are included for the visualization of reports and to aid decision-making through data.

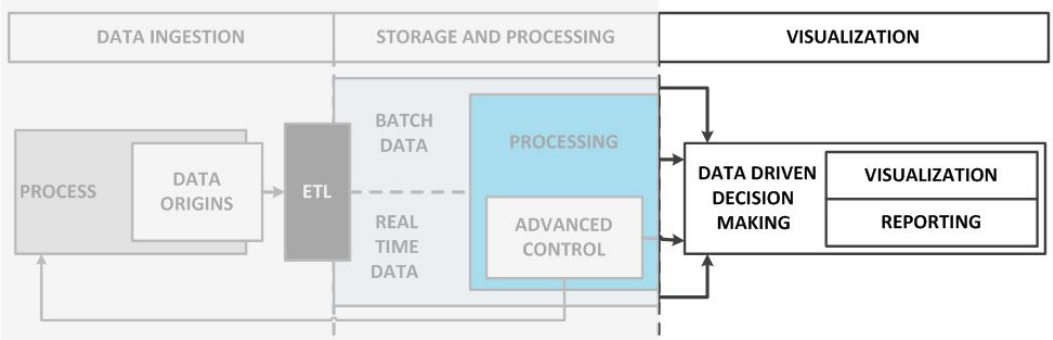


FIGURE 6.21: Data visualization in the data architecture.

These applications can use historical data, real-time data or application results. They can be specific applications or be part of a business planning system, ERP.

Data architecture integration with Fábrica Vitoria systems

Figure 6.22 specifies the summary of the architecture proposed as a solution applicable to the paint plant of the Vitoria factory. This figure shows the data generation and synchronization procedure for each data source presented in Figure **??** and is applicable to the set of painting processes (TTS, KTL, Sealing, Primer, Coating, Finish, Waxing).

6.5.3 Data synchronization proposal at the Vitoria Factory

The objective of data synchronization is to generate, using the previously proposed methodology, a data board in which each line contains the information associated with a bodywork. To achieve this objective, a series of requirements are necessary:

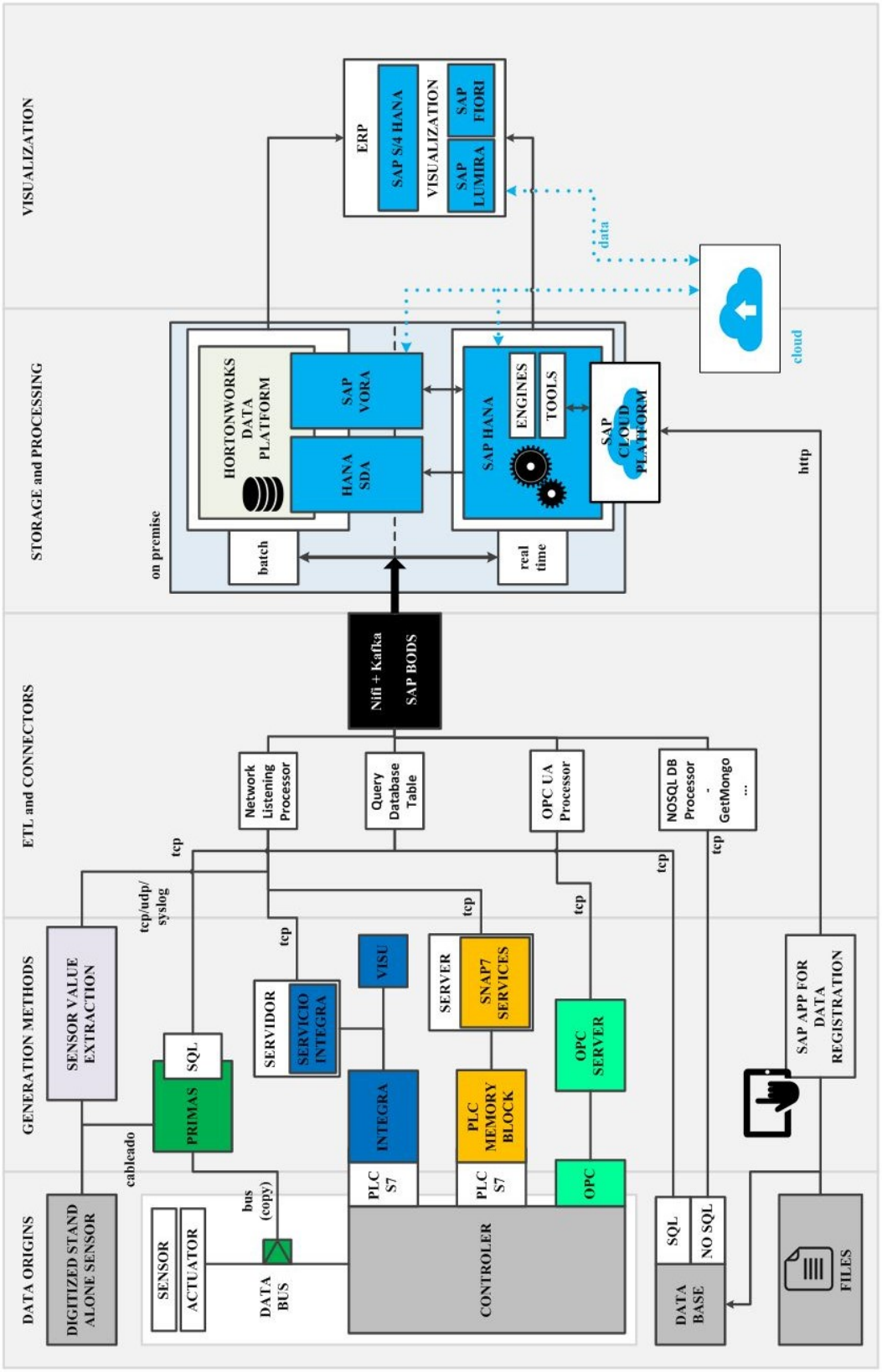


FIGURE 6.22: Proposed Generation Architecture for Fábrica Vitoria.

- Time synchronization. Every paint shop equipment share the same Timestamp. It is recommended that this Timestamp is a single value, for example,

the number of milliseconds since a specific date. This is justified by the mistakes in the interpretation of dates depending on the language in which they are expressed. As an example, during the development of the case study, there were problems with the data acquisition dates since the Timestamps alternated the date format in English (MM / DD) and Spanish (DD / MM).

- Spatial synchrony. The vehicle body can be located at every step of the paint shop process. This is very important to accurately relate the value of the variable to the moment when it was relevant for the paint film. This requirement is the result of a lesson learned during the case study since, in the coating process, only the getting in and going out of the process timestamps are captured, so that the relevance times of some variables cannot be precisely adjusted to the moment or action in which they were relevant.
- Temporal determinism. The starting and ending times of the operations carried out in the processes are known, that is, the start and stop times of the robots are known, opening and closing times of needles and so on. Meeting this requirement improves the accuracy of the data set. Figure 6.23 exemplifies this temporal determinism.

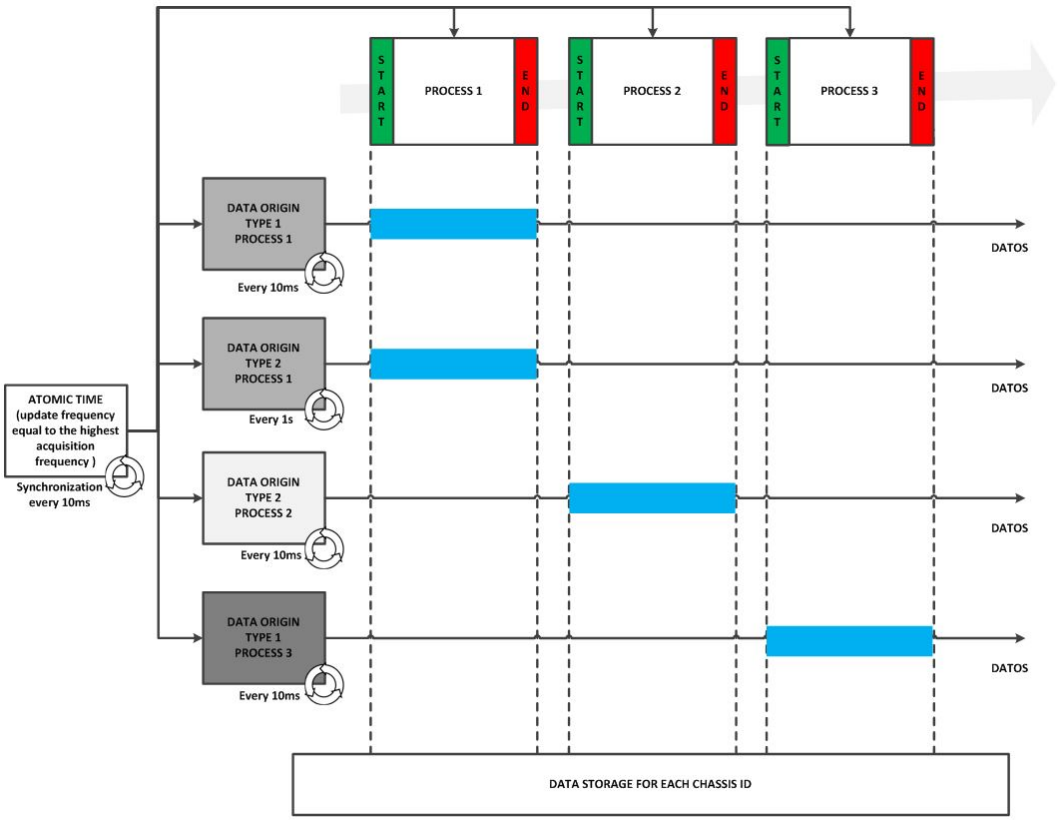


FIGURE 6.23: Temporal determinism example.

An example of variable synchronization for a unique ID chassis is described in Figure 6.23 from the perspective of temporal determinism. There are a number of processes (process 1, process 2 and process 3), each with the corresponding data sources. These data sources are continuously releasing data, but the values of the variable are stored only when the variable is relevant for the quality of the paint film (shown by the segment painted in blue). This is defined as the moment when

the variable "touches the chassis". This concept is key to define the synchronization methodology, since it is necessary to know where the body is in the process (spatial synchrony), the starting and ending Timestamps of each process, as well as the Timestamps of each variable provided by each data source (time synchronization). This time synchronization can be specified for each small action within each process, that is, each process is made up of a sequence of actions. The precision of the data will be the better the more precise is the division of the process in its sequence of actions. This is exemplified in Figure 6.24.

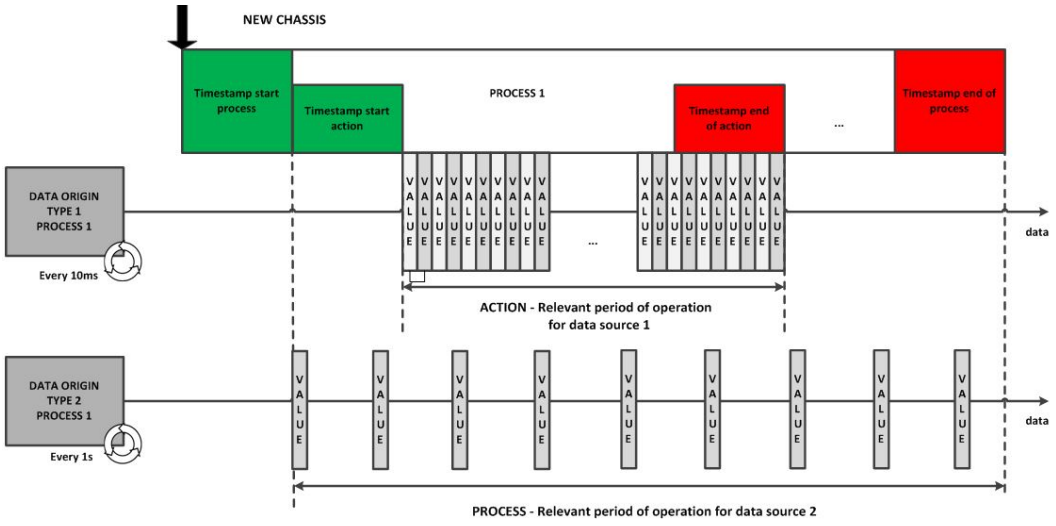


FIGURE 6.24: Action synchronization in the painting process.

As shown in Figure 6.24, using the timestamps of the actions within the process can be more precise as the time of relevance of each variable is known and, in general, each variable influences the process for a short period. Collecting data only when the variable is relevant for the paint shop process is the optimal solution for developing a predictive model of paint film quality. With other solutions, such as, predictive maintenance or solutions not directly related to the product, the data must be synchronized with a specific procedure.

6.5.4 Data generation and synchronization procedure generalization

Here, as a resume of this section, the actions required by the generation proposal are ordered and presented.

First, the work flow of the paint shop is analysed and all the data sources are listed. This is done by differentiating the following categories of data sources:

- Non digitized data sources.
- Digitized and non usable data sources.
- Digitized and usable data sources.

For the first case, Non digitized data sources, the generation procedure has been defined in the Data extraction proposals section distinguishing the procedures for unmeasured variables, the variables measured with purely mechanical instruments and the variables that are recorded on paper record sheets. In the second case, Digitized and non-exploitable data sources, referring to data sources from which data

could not be generated. Here, a generation proposal is made by applying an ordered procedure of actions to generate the data. The actions are ordered from least to greatest difficulty of implementation.

- Create an application to introduce and upload the parameters registered in the mixing room to SAP (avoid the use of Excel files).
- Carry out generation tests with the OPC method in the test cell of the paint shop and analyse the viability of installing an OPC connection in the data source systems that are of interest.
- Apply the specific methods for each data source (PRIMAS, Integra, OPC).

In the third case, Digitized and exploitable data sources, refers to the data sources used from which usable data have been obtained. As a form of integration of these data origins in the Fábrica Vitoria systems:

- Integrate the data from the GI and GP databases into the SAP plant system.
- Import the files from the Q-EYE quality tunnel to the system.
- Import the file generated by the paint booth equipment into the SAP system (downloaded files generated by the flow control equipment).

Considering the specific generation methods that have been analysed from the State of the Art in generation in the Sindelfingen paint shop, the appropriate method is assigned to each data source considered digitized not exploitable:

- For the stand alone sensor, a Sensory Capture Infrastructure method (such as a datalogger system) or PRIMAS is recommended.
- For the files generated by systems such as, for example, artificial vision systems, a copy and upload process to a plant database is generated and thus their data can be imported from it by means of a database query (query).
- To generate data from Machine Controller-Visu systems such as DÜRR machines, there are several options depending on the data source, the generation points of each source or the desired data acquisition frequency. For instance:
  - PRIMAS for real-time data from the data bus.
  - INTEGRA for VISU data.
  - OPC for systems with this OPC interface capability.
  - SNAP7 for Siemens S7 controllers.

The following section presents the procedure for selecting the most appropriate specific generation method for the latter case (Machine Controller-VISU).

Machine data generation method selection procedure

In this section, a procedure for the selection of the process data generation method is presented. This procedure depends on the data source from which values have to be obtained. For some data sources, this method selection procedure is not applicable as the only generation alternative is to apply a specific method. Specifically:

- To generate VISU data, the Integra method is used.
- To generate data that are saved in files, the file is imported into the database.

- To generate data that is in a database, it is done through a query procedure to said database.

In the case of generating machine data, it is the one that offers the greatest variability, since the frequency with which we want to acquire the data or the point of generation of the data of the variable of interest influences. Therefore, Figure 6.25 shows the proposed procedure for generating machine data in which there is an associated controller and with which it has communication. The digitized stand alone sensor case is also included.

Following the flow diagram in Figure 6.25, first analyse if the translations of the data strings that are generated by process machines or the measurement scales of the sensors are available. That is, if the variable codification of the machine can be decoded so that the acquisition and transformation is carried out in a transparent way. In case of not having the decodification of the variables, that is, not being able to identify the information generated by the machine, the necessary actions must be taken to obtain that information, either through the supplier of the machine or through a third party. The next point to be considered is the acquisition frequency required for each variable. Lastly, the data generation methods depend on the data generation point.

- In the case of real time data acquisition (acquisition frequency greater than the PLC cycle time), the analysis of the state of the art in data generation in Sindelfingen recommends the PRIMAS system.
- If the acquisition frequency is less than the cycle time, the acquisition point must be considered. If the variable is in the memory of a non-Siemens PLC, the OPC method is used. If the controller is Siemens, the use of the SNAP7 method or the OPC method is recommended.

To recap, it should be noted that the SNAP7 method is free but where the acquisition times, verified by expert knowledge of the data team at Sindelfingen, are approximately 1 second. In the case of OPC, it depends on the configuration of the system, so it can be said that it is not deterministic and in each case data will be obtained at a different frequency. For its part, Siemens provides an online tool on its platform to approximate the acquisition time of the OPC system for a series of configuration combinations [16]. Although it can be observed (table of cases) that the acquisition times are generally shorter with the OPC method, it must be considered that the use of this method entails the cost of using the OPC server license.

Figure 6.26 shows an example of S7 PLC data acquisition times through OPC UA offered by the Siemens tool.

This example shows, for a simple network configuration, the times of writing, reading and transmission of data blocks from PC to PLC and from PLC to PC as well as the monitoring of variables. It is observed that the average time for data reading is 30.5 milliseconds, which for some applications may not be a sufficiently fast acquisition frequency.

Process data synchronization procedure

Next, Figure 6.27 presents the proposed methodology for the synchronization of process data with the bodywork as the primary key.

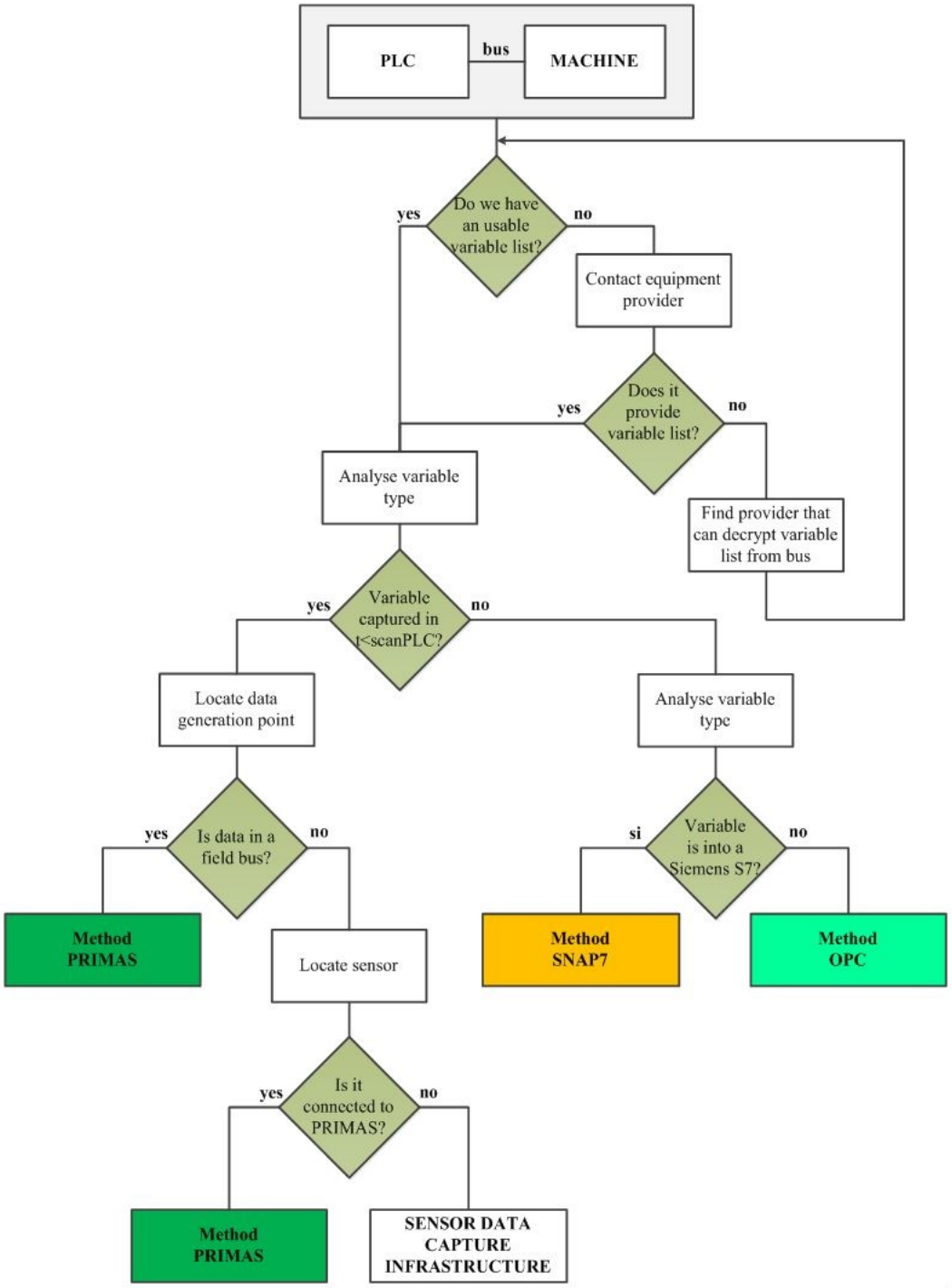


FIGURE 6.25: PLC-Machine data generation procedure.

As can be seen in the flow diagram that represents the procedure,the relevant points are the presence of a new chassis ID at the beginning of the process and the time synchronization between equipment. Once these steps have been achieved, the synchronization is based on ordering the Timestamps of the variables as they appear in their corresponding action or process. That is, once we have the process start Timestamp, knowing the variables that belong to that process, the values corresponding to the Timestamps after the process start Timestamp (or action) and before the process

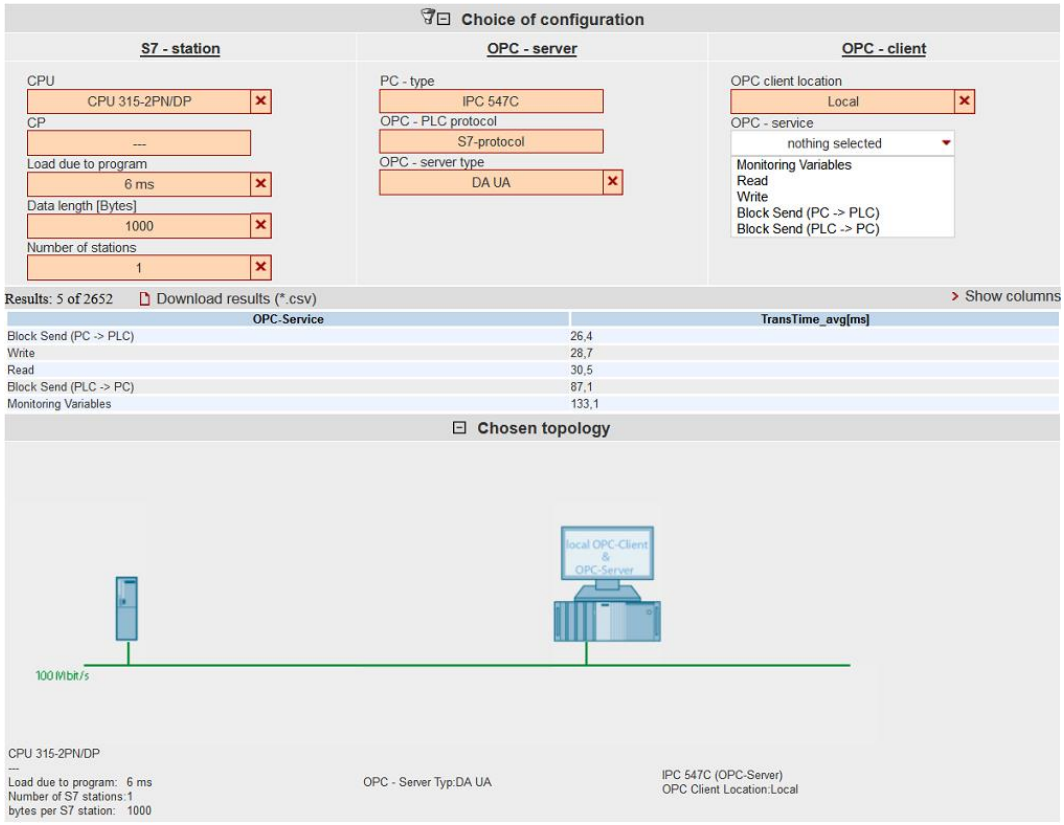


FIGURE 6.26: Execution times in connections with S7 PLC via OPC UA.

(or action) stop Timestamp are assigned to the bodywork. These actions are carried out throughout the body painting sequence until the corresponding data set line is completed.

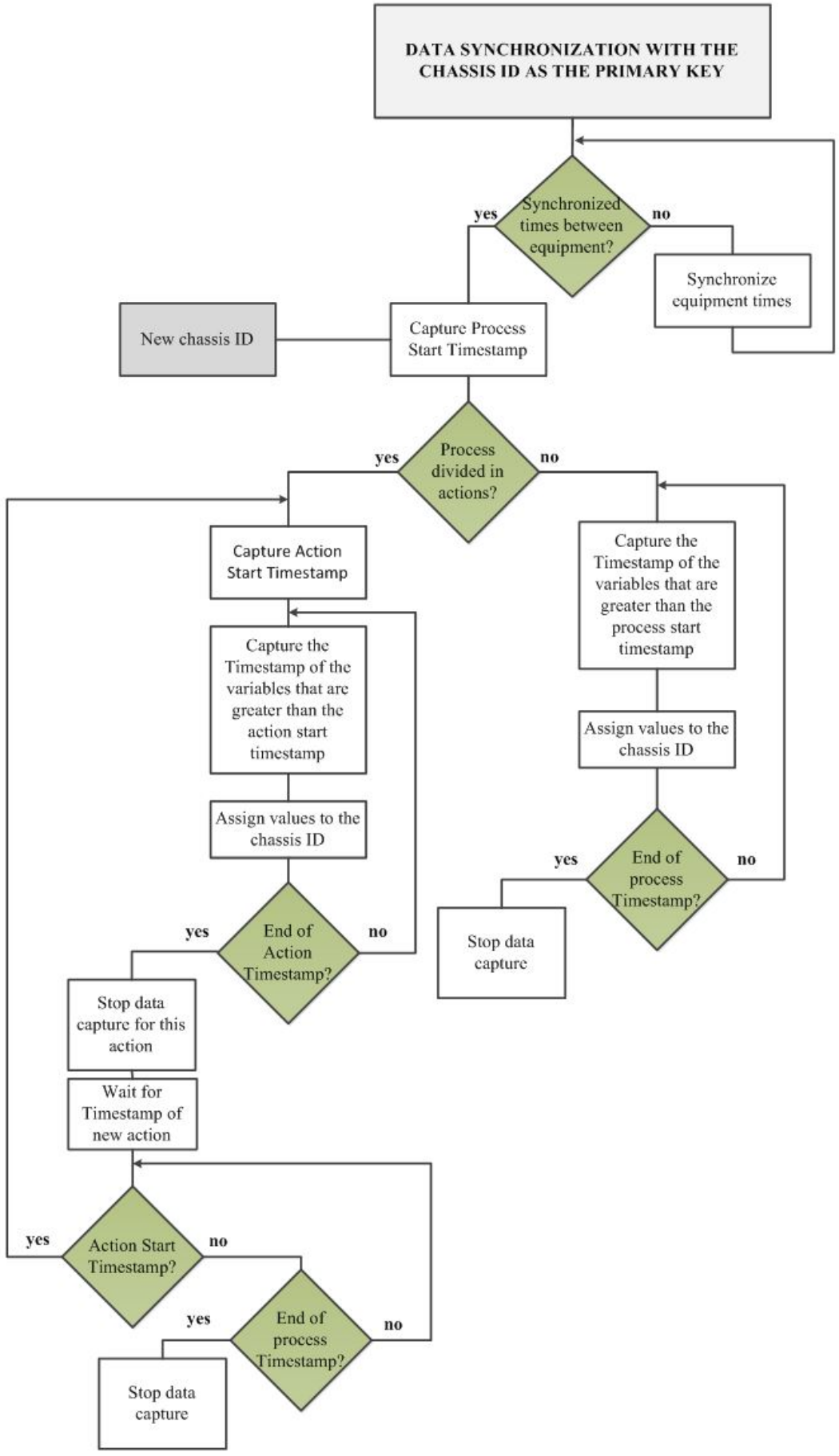


FIGURE 6.27: Process data synchronization procedure.

CHAPTER 7

Conclusions

7.1 Main conclusions

Related to objectives:

- Objective 1: analyse the state of extraction of variables from the factory.

  The plant has been toured in such a way that the relevant variables of each subprocess have been identified and the possible origins of each data have been searched. The result has been reflected in Paint shop case study: Mercedes Benz Fábrica Vitoria (Chapter 6). Only the 18% of the identified variables were digitized and their value was accessible. Regarding the extraction of process data, it is concluded that there is no single recipe as a generation method. Each of the methods has its specific application case regarding the capacity to read and / or write data (volume and frequency of data acquisition and transformation), the industrial equipment from which they are capable of extracting data (compatibility) and they can be a proprietary or free data extraction system (with acquisition costs, technical support and development difficulty as variables to take into account).

- Objective 2: define a set of output variables that can best describe quality and that can be measured.

  The paint shop at the Mercedes Benz factory in Vitoria has an automated system for assessing the quality of the paint layer of the finished vans at the end of the painting process. This evaluation, which is carried out mechanically, has a great consistency and repeatability since the detection criteria is constant.It is performed on all the vans that leave the line, which facilitates the location and selection of the variables that indicate quality. This automated defect detection system can characterize defects among several scenarios, with dirt and crater being the most frequently detected defects. In order to increase the number of samples to feed the dataset (in this case painted vans with or without defects), these two defects have been selected as the output variable that can best describe quality. If they are detected, they are indicators of non-quality and if they are not detected, it is considered that the painting process has been carried out with quality.

- Objective 3: for this or these output variables, generate a set of input variables that can explain the outcome of the output variables.

  The proposed methodologies have been applied to define and generate a data set with the values of process variables that are considered the inputs of those dirt and crater variables as they evaluate quality (outputs). The process has been traversed identifying the variables that were relevant in each operation of the paint shop process. Each variable has been univocally identified. A search was carried out for any possible data source for these variables, discarding those that did not meet the necessary criteria. The synchronization of input and output data was carried out using each unique chassis ID as the primary synchronization key, taking into account, through the use of process Timestamps, the value of the input variable only at the moment when it "touched" the chassis.

- Objective 4: describe a methodology to develop a predictive model of the paint film quality of the car bodies.

  Once the data set that relates input and output has been generated, before using it to train the predictive model, a data quality analysis has been performed, discarding for model training those chassis (samples) that did not meet the required data quality, either by having out-of-range values, incomplete data or others.It has also been evaluated that the quality obtained by the samples was not shifted towards an over representation of good or bad quality, to avoid model overfitting. The SMOTE technique has been used to generate synthetic samples and improve the model results. Supervised machine learning algorithms have been implemented (the best performing algorithms were decision tree in the first and second round and simple logistic regression in the third round, the number of samples increased in each round as more samples were obtained from the paint shop). The model evaluation criteria were accuracy and AUROC. The values obtained were satisfactory, so it is concluded that there is a correlation between the values of the input variables and the paint film quality obtained.

- Objective 5: apply and test these methodologies in a real paint shop.

  The methodologies have been applied to the Mercedes Benz Fábrica Vitoria paint shop and the results are presented and discussed in Paint shop case study: Mercedes Benz Fábrica Vitoria (Chapter 6).

As conclusions drawn from the study of the state of the art in data generation at the Sindelfingen paint shop, the following should be noted:

- Regarding the generation of process data, it is concluded that:
  - The generation structure in Fábrica Vitoria is similar to the one studied in Sindelfingen, with the same sub-processes to which the chassis surface is subjected during the painting process and, in addition, they have equivalent machinery. Due to this similarity of process and generation cases, the proposed generation methodology follows the steps carried out in the German factory.
  - The architecture proposal that supports data management includes a structure based on Hadoop technologies but integrated into the SAP solution deployed at Fábrica Vitoria.

  - The proposed generation methodology, rescues the methodology used in the development of the proof of concept and proposes the adoption of generation methods used within the Daimler group as mentioned in point a.

- Regarding the process data synchronization, it is concluded that:
  - A data synchronization proposal has been made that includes the following requirements:
    * It is necessary to locate the bodywork in the process at all times.
    * Time synchronization of systems is necessary. The synchronization frequency has to be at least the same as the highest data acquisition frequency.
  - Following the flow of actions proposed for synchronization, it is possible to fill in the data line corresponding to each chassis number so that this is the primary key. This form of synchronization is suitable for successfully developing advanced analytic applications based on product qualities.

The implementation of the proposed methodology will allow the development of predictive models of product and process quality as well as the development of digital twins.

7.2 Publications

J. Salcedo-Hernández, J. García-Barruetabeña, I. Pastor-López and B. Sanz-Urquijo, "Predicting Enamel Layer Defects in an Automotive Paint Shop," *IEEE Access*, vol. 8, pp. 22748-22757, 2020, doi: 10.1109/ACCESS.2020.2969816.

7.3 Recommendations for future work

The following suggestions for future work are made in order to improve the digitization possibilities of industrial companies.

- Continue working on the digitization of the paint process, making more variables available to generate the paint shop dataset and thus continue to increase the accuracy of the model until the time comes when it is possible to switch from a corrective to a preventive production mode.
- Investigate the application of edge computing to generate corrective actions near the process that are based on predictions of the process behaviour.
- Work on a non-proprietary solution to extract machine data, for example from a communication bus, as the lack of data extraction methods, especially on legacy devices (computing device or equipment that is outdated, obsolete or no longer in production) makes it difficult to generate industrial datasets.

Bibliography

Bauernhansl, Thomas (2017). "Die vierte industrielle Revolution–Der Weg in ein wertschaffendes Produktionsparadigma". In: *Handbuch Industrie 4.0 Bd. 4*. Springer, pp. 1–31.

Bley, Katja, Christian Leyh, and Thomas Schäffer (2016). "Digitization of German Enterprises in the Production Sector-Do they know how "digitized" they are?" In:

Chawla, Nitesh V et al. (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

Chenhall, Robert H (1997). "Reliance on manufacturing performance measures, total quality management and organizational performance". In: *Management Accounting Research* 8.2, pp. 187–206.

Cole, GS and AM Sherman (1995). "Light weight materials for automotive applications". In: *Materials characterization* 35.1, pp. 3–9.

Fawcett, Tom (2006). "An introduction to ROC analysis". In: *Pattern recognition letters* 27.8, pp. 861–874.

Garavaglia, Matteo et al. (2019). "Process development and coaxial sensing in fiber laser welding of 5754 Al-alloy". In:

Garner, Stephen R et al. (1995). "Weka: The waikato environment for knowledge analysis". In: *Proceedings of the New Zealand computer science research students conference*, pp. 57–64.

Business intelligence and recording of process data at GM paint shops (2016). Strategies in Car Body Painting. Berlin.

Hall, Mark et al. (2009). "The WEKA data mining software: an update". In: *ACM SIGKDD explorations newsletter* 11.1, pp. 10–18.

Hermann, Mario, Tobias Pentek, and Boris Otto (Jan. 2015). "Design Principles for Industrie 4.0 Scenarios: A Literature Review". In: DOI: 10.13140/RG.2.2.29269.22248.

Kuziak, R, Rudolf Kawalla, and Sebastian Waengler (2008). "Advanced high strength steels for automotive industry". In: *Archives of civil and mechanical engineering* 8.2, pp. 103–117.

Marzuki, Mohammad Al Bukhari, Mohammad Firdaus Mohammed Azmi, and Rafidah Laili Jaswadi. "Design optimisation of automotive component through numerical investigation for additive manufacturing". In: ().

Materials, Best (2019). *Sheet metal gauge chart Gauge to thickness chart*. https://www.bestmaterials.com/PDF_Files/sheet-metal-gauge-chart.pdf. Accessed: 2019-07-22.

Mönch, Lars, Jens Zimmermann, and Peter Otto (2006). "Machine learning techniques for scheduling jobs with incompatible families and unequal ready times on parallel batch machines". In: *Engineering Applications of Artificial Intelligence* 19.3, pp. 235–245.

Paint Defects. URL: https://uk.ppgrefinish.com/en/paint-defects/.

Peng, Ying, Ming Dong, and Ming Jian Zuo (2010). "Current status of machine prognostics in condition-based maintenance: a review". In: *The International Journal of Advanced Manufacturing Technology* 50.1, pp. 297–313. ISSN: 1433-3015. DOI: 10.1007/s00170-009-2482-0. URL: https://doi.org/10.1007/s00170-009-2482-0.

Pipino, Leo L, Yang W Lee, and Richard Y Wang (2002). "Data quality assessment". In: *Communications of the ACM* 45.4, pp. 211–218.

Press, Gil (2016). *Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says*. URL: https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/.

Rambabua, D. et al. (2018). "Advanced product design and optimization using sap plm and integration with supply chain processes in automotive manufacturing". In: *Journal of Web Engineering* 17.6, pp. 2089–2103.

Steyerberg, Ewout W et al. (2003). "Internal and external validation of predictive models: a simulation study of bias and precision in small samples". In: *Journal of clinical epidemiology* 56.5, pp. 441–447.

Streitberger, Hans-Joachim and Karl-Friedrich Dossel (2008). *Automotive Paints and Coatings*. English. 2. Aufl. US: Wiley-VCH, pp. 89–127. ISBN: 9783527309719. URL: http://ebooks.ciando.com/book/index.cfm/bok_id/504868.

Taleb, Nassim Nicholas (2007). *The Black Swan: The Impact of the Highly Improbable*. Random House Group. ISBN: 1400063515.

Tornero, Josep et al. (Mar. 2012). "Detección de Defectos en Carrocerías de Vehículos Basado en Visión Artificial: Diseño e Implantación". In: *Revista Iberoamericana de Automática e Informática Industrial RIAI* 9, 93–104. DOI: 10.1016/j.riai.2011.11.010.

Wang, L. and J. Qi (2009). "The Real-Time Networked Data Gathering Systems Based on EtherCAT". In: *2009 International Conference on Environmental Science and Information Application Technology*. Vol. 3, pp. 513–515. DOI: 10.1109/ESIAT.2009.489.

Wollschlaeger, Martin, Thilo Sauter, and Juergen Jasperneite (2017). "The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0". In: *IEEE Industrial Electronics Magazine* 11.1, pp. 17–27.

Wolpert, David H (1996). "The lack of a priori distinctions between learning algorithms". In: *Neural computation* 8.7, pp. 1341–1390.

Wolpert, David H, William G Macready, et al. (1997). "No free lunch theorems for optimization". In: *IEEE transactions on evolutionary computation* 1.1, pp. 67–82.

Zhu, C. et al. (2015). "A Tree-Cluster-Based Data-Gathering Algorithm for Industrial WSNs With a Mobile Sink". In: *IEEE Access* 3, pp. 381–396. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2015.2424452.