QU4LITY

DIGITAL MANUFACTURING PLATFORMS FOR CONNECTED SMART FACTORIES

D3.5 Big Data and Analytics Infrastructure

Deliverable Id :	**D3.5**
Deliverable Name :	**Big Data and Analytics Infrastructure**
Status :	**FINAL**
Dissemination Level :	**PU**
Due date of deliverable :	**31/12/2019**
Actual submission date :	**30/12/2019**
Work Package :	**WP3**
Organization name of lead contractor for this deliverable :	**Athens Information Technology (AIT)**
Author(s):	**John Soldatos, Nikos Kefalakis**
Partner(s) contributing:	**FAGOR, IKER, ENG, LKS, ATLAS, FHG, VTT, TNO, EPFL, TUDO**

Abstract: This deliverable is devoted to presentation of Big Data enablers for Zero Defect Manufacturing (ZDM). The enablers are structured in two main parts, namely Big Data infrastructures (including Big Data databases and data processing middleware) and Data analytics algorithms (including data mining and machine technique for ZDM.

Co-funded by the Horizon 2020 Programme of the European Union

www.QU4LITY-project.eu

Contents

HISTORY

Version	Date	Modification reason	Modified by
0.1	12/07/2019	Initial Table of Contents to be shared & discussed with WP leader	AIT
0.12	11/09/2019	First Draft of the Document with the Introduction filled-in and more detailed placeholders for partners' contributions	AIT
0.13	11/09/2019	Initial Inputs in Platforms and Algorithms	ATLAS, VTT
0.14	13/09/2019	Further refinements to the deliverable structure	AIT
0.15	11/10/2019	Inputs from FHG-IGD	FHG
0.16	11/10/2019	Inputs from ATLAS on RUL Prediction	ATLAS
0.17	25/10/2019	Description of AIT's QUARMA algorithms and inputs on DL based RUL prediction	AIT
0.18	28/10/2019	Inputs on Sensor Deployment optimization from TNO	TNO
0.19	28/10/2019	Description of the relevant of the Big Data platforms to the QU4LITY RA	AIT
0.20	29/10/2019	Inputs regarding the FAR-EDGE DDA platform and its use in QU4LITY	AIT
0.21	12/10/2019	Inputs on AITs DDA platform	AIT
0.23	25/11/2019	Updated Contribution from ATLAS – Merging of Data Driven and Model Driven Approaches	ATLAS
0.24	26/11/2019	Inclusion of a Discussion of Platforms Alignment to the QU4LITY RA	AIT
0.25	03/12/2019	Updated input and visuals from FHG-IGD	FHG-IGD
0.30	05/12/2019	Various Edits, Updates and Quality Improvements of the Document	AIT
0.32	09/12/2019	Inputs from TUDO	TUDO
0.35	10/12/2019	Various edits and formatting	AIT
0.36	24/12/2019	Internal deliverable review	JSI
0.50	26/12/2019	Addressing comments from Quality Review	AIT
0.55	27/12/2019	Fine Tuning and Formatting	AIT
1.0	27/12/2019	Preparation of Version for Delivery to EC	AIT

1. Introduction

1.1 Scope and Purpose of the Deliverable

QU4LITY is developing and deploying solutions for the next generation of Zero Defect Manufacturing (ZDM) capabilities in the Industry4.0 era. The project's solutions will therefore leverage Cyber Physical Systems (CPS) and advanced digital technologies (e.g., Big Data, Edge/Fog Computing, Artificial Intelligence). QU4LITY is concerned not only with demonstrating Industry4.0 ZDM solution in productions lines, but also with providing reusable building blocks for developing and integrating such solutions. To this end, the third workpackage (WP3) of the project is developing various digital enablers and building blocks that can support the development of intelligent, automated and secure ZDM solutions. These digital enablers include components that support Big Data collection, storage, management and processing, given that the vast majority of ZDM applications are and will be data-intensive. In this context, the present deliverable is devoted to the presentation of the Big Data building blocks that will be developed and used in the scope of the project. These building blocks will support the development, deployment and operation of the data-intensive pilot systems of the project, while at the same time serving as prominent and characteristic examples of BigData for ZDM. Some of these building blocks will be also integrated and promoted through the project's market platform, as part of the project's efforts to provide the European manufacturing community with concrete examples of Big Data solutions and with demonstrations of their use for quality management and ZDM.

The Big Data digital enablers and building blocks of the deliverable are classified in two broad categories:

- **Big Data Platforms for Quality Management and ZDM**: This category includes distributed file systems, distributed databases, streaming middleware platforms, data collection platforms, visualization platforms and other platforms that deal with large amounts of data that possess the Big Data V's (i.e. Volume, Variety, Veracity, Velocity) and which can be used to support ZDM and Quality Management applications. In particular, the deliverable presents a catalogue of such platforms and their functionalities.
- **Data Analytics Algorithms for Quality Management and ZDM**: This category comprises a set of algorithms and techniques for analyzing production data (e.g., quality data, machine status data, data from business information systems) towards deriving insights for improving ZDM and achieving the vision of automated and intelligent quality management in the factories of the future. Hence, the deliverable presents a catalogue of data mining techniques, including Machine Learning (ML) and Deep Learning (DL) techniques, which are fit for purpose for ZDM and quality management in future manufacturing.

The platforms and algorithms that are presented in this deliverable are mostly based on partners' background IP (Intellectual Property), notably IP that has been developed in past projects and initiatives relating to quality management, predictive maintenance, industrial automation and other Industry4.0 related activities. These background components will be advanced and customized as a means of becoming tailored to ZDM and quality management requirements. The advancements to take place range from their (re)configuration in-line with ZDM requirements to enhancements to their core functions (e.g., statistical processing techniques used). Moreover, several advancement and customizations of background components of the partners are destined to support the unique value propositions of the QU4LITY project i.e. the support of an autonomous and cognitive quality management paradigm.

Note also that the presented Big Data platforms and algorithms for ZDM are not exhaustive, as many other platforms and algorithmic enablers have been proposed in the literature and are used in ZDM projects. Nevertheless, the enablers that are reported in this deliverable are representative of the capabilities of Big Data technologies and their merits for effective quality management in manufacturing environments. Thus, they could certainly help the Industry4.0/ZDM community understand and leverage Big Data technologies in their quality management processes.

It should be noted that this deliverable is a comprehensive report on the Big Data enablers of the QU4LITY project. A subsequent version of this deliverable (i.e. deliverable D3.6) will be accompanied with the final prototype implementation of these enablers, including the planned enhancements needed to support the QU4LITY paradigm. In the scope of deliverable D3.6, the QU4LITY partners will report any changes/updates in the components and platforms that are presented in this deliverable, including changes and updates triggered by the practical use of the enablers in the project's pilots.

1.2 Methodology

As already outlined, the starting point for reporting the Big Data enablers (i.e. platforms and algorithms) of the QU4LITY project is the set of Big Data related IP of the partners that contribute to this deliverable. Nevertheless, these background technologies will be enhanced and adapted to QU4LITY requirements, including both business and technical requirements. To this end, the design and the development of the Big Data enablers of the project is driven by stakeholders' requirements regarding autonomous quality in Industry4.0 manufacturing, as well as by the ZDM Reference Architecture (RA) of the project. The latter requirements and RA have been both developed in WP2 of the project and there are documented in distinct deliverables. These deliverables have therefore been considered in developing this deliverable.

Likewise, the development of the Big Data enablers is also driven by the QU4LITY pilots, given that all of them will be deployed, validated and evaluated in one or more pilots in WP7 of the project. Thus, the partners have considered pilot requirements based on relevant discussions and meetings at the pilot sites with the representatives of the pilot sites. Obviously, these meeting involved the partner/provider of each enabler and representatives of the pilot(s) where the enabler is to be deployed and used.

From a development standpoint, an agile, iterative approach is followed, based on two broad developments and reporting cycles. An initial implementation of each enabler (including adaptation to QU4LITY requirements) will be performed as part of this deliverable based on the pre-existing knowhow of the partners and initial requirements of the pilots. A second and final implementation cycle will then lead to the final version of each enabler. This second cycle can be broken down in smaller sub-cycles, each one related to the progressive advancement of each enabler and its improvement based on feedback from the lab validation and from the pilot deployment of the enabler in pilot production lines (WP6/WP7).

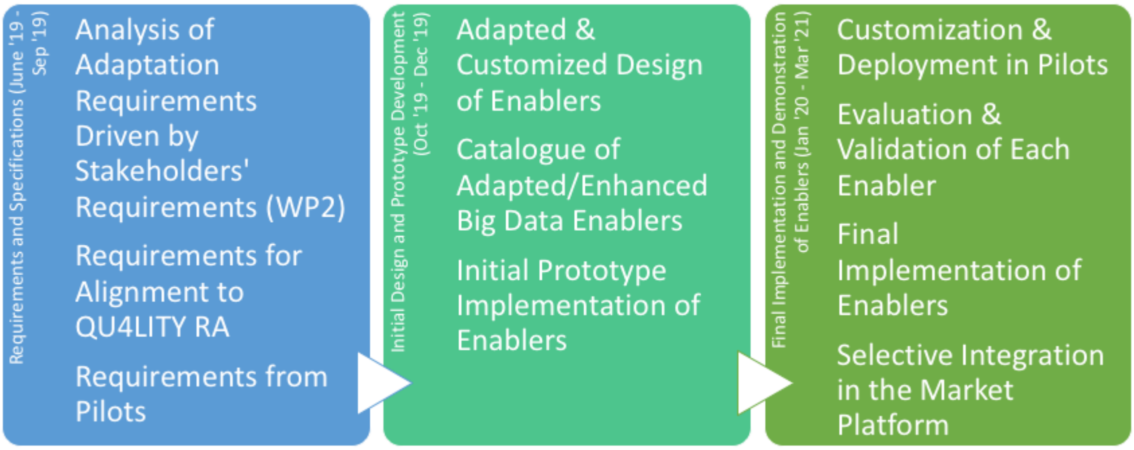


Figure 1: Overview of the Methodology Used for Producing this Deliverables and Planning for Delivery of the Final Version of the Big Data Enablers

Overall, the methodology for the development and delivery of the Big Data enablers of the project comprises the following phases (as also depicted in Figure 1):

- **Phase 1 – Requirements and Detailed Specifications for the Customization & Enhancement of Big Data Enablers**: During this phase the partners identified the ways in which the various background technologies (enablers) must be enhanced and customized to the needs of the QU4LITY project. To this end, requirements from WP2, the QU4LITY RA and requirements from the pilots have been considered.
- **Phase 2 – Design and Initial Prototype Implementation of Enhanced Enablers**: During this phase the design of the customized and/or enhanced enablers was produced and documented in this report. Most of the enablers have also an initial prototype implementation that supports the QU4LITY autonomous and cognitive quality paradigm.
- **Phase 3 – Final Implementation of Enhanced & Customized Enablers**: This is the final phase of the project's methodology that will lead to the final version of the enhanced & customized Big Data enablers. Development activities in this phase will consider feedback from the deployment, validation and evaluation of

the enablers, both in lab environments and in the pilot production lines that support the pilot's activities of the project.

1.3 Relation to Other Deliverables

Based on the above-outlined introductory statements and methodology, one can infer the close linking of the present deliverable with the following documents/deliverables of the project:

- **D2.1 Analysis of User Stories and Stakeholders' Requirements (Version 1),** which was one of the main sources for ZDM requirements that all providers of Big Data enablers have been considered in customizing/enhancing their components.
- **D2.3 Autonomous Quality Vision for ZDM and Quality Management Excellence (Version 1),** which was another source of requirements that drove the customization and enhancement of the Big Data enablers.
- **D2.11 Reference Architecture and Blueprints (Version 1),** given that the platforms and algorithms of this deliverable can be deployed in-line with the RA of the project and in-line with relevant solution blueprints specified in deliverable D2.11.
- **D3.3 HPC and Cloud Resources for ZDM (Version 1),** given that some of the platforms that are described in this deliverable are cloud-based.
- **D3.7 Fog Nodes and Edge Gateways for ZDM deployments (Version 1),** since some of the algorithmic enablers of the project will be deployed in fog/edge nodes, while some of the platforms that are outlined in this deliverable adhere to the edge/fog computing paradigm.
- **D7.1 Detailed Pilot Specification and Report on Pilot Sites Preparation (Version 1),** which describes the current status of the pilots' preparation along with pilot requirements that have been taken into account in properly adapting the Big Data enablers of this deliverable.

There are more deliverables that are related to the present one, but the ones listed above have the most prominent linking.

1.4 Deliverable Structure

The rest of the deliverable is structured as follows:

- Section 2 (following this introduction), provides a brief overview of the requirements that have driven the Big Data enablers of the deliverable. Note that it emphasizes on high-level requirements that are applicable to most of the enablers, rather than low-level requirements pertaining to specific enablers. It also presents some structuring principles that are associated with the deployment and use of the various enablers in ZDM systems. These structuring principles are aligned to the project's RA for autonomous ZDM systems and deployments.
- Section 3 provides the catalogue of Big Data platforms, including detailed information about the design, the prototype implementation and the use of each one.

- Section 4 is devoted to presenting the catalogue of analytics algorithms that will be used in the project, including an analysis of their characteristics that make them suitable for ZDM.
- Section 5 is the final and concluding section of the paper.

2. Driving Requirements

2.1 QU4LITY Architecture and Platform Development Requirements

The QU4LITY Big Data solutions are integral elements of the QU4LITY ZDM solution, as they are part of the QU4LITY RA. This is depicted in Figure 2, which illustrates different components and views of the RA.

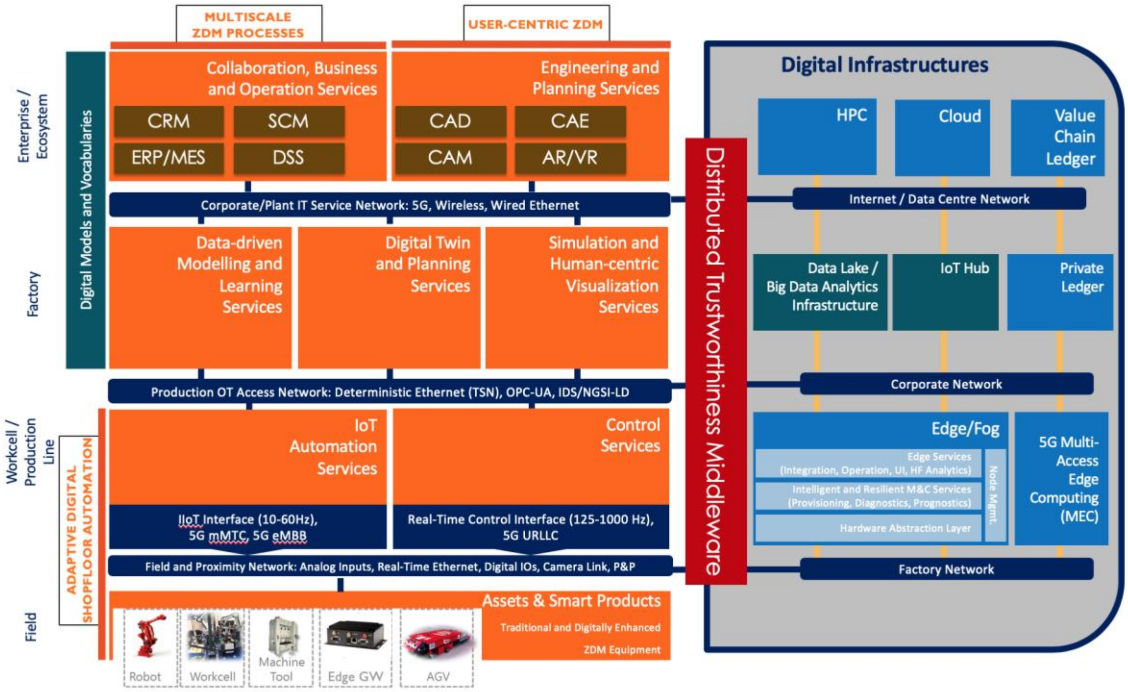


Figure 2: Snapshot of the initial version of the QU4LITY RA

In-line with deliverable D2.11, the Big Data infrastructure is part of the deployment and implementation views of the architecture, which comprise the following components:

- **The Data Lake and Big Data Analytics Infrastructure**, where structured, semi-structured and unstructured data are stored and managed, in scalable and resilient ways. Big Data Analytics infrastructures operate over data lakes in order to analyze data and visualize quality management related insights based on proper dashboards and visualizations. The QU4LITY platforms that are presented in this deliverable comprise data storage infrastructures in the form of data warehouses, databases and data lakes, while at the same time provide enabling the execution of ML algorithms such as deep learning.
- **Digital Models and Vocabularies**, where a set of sharing digital models and vocabularies reside. These digital models provide the means for exchanging information across the diverse components of a QU4LITY solution. In practice, they provide a basis for syntactic and semantic interoperability. In the present deliverable, each QU4LITY Big Data platform is presented and used in isolation

from the others. However, the use of QU4LITY digital models that are specified in other WP2/WP3 deliverables can enable the interoperability of these different platforms and can therefore enable their simultaneous deployment as part of a QU4LITY compliant solution.

- **IoT Hub**, which provide a central communication hub between IoT devices and applications managing them. It supports bidirectional communications both from the device to the cloud and from the cloud to the device. Note that several of the data collection and data management platforms that are presented in this deliverable provide the means for accessing data from IoT devices and therefore belong in the IoT Hub component of the QU4LITY RA.

Overall, the platforms and algorithms that are presented in this deliverable are positioned in the Big Data Infrastructure and IoT Hub components of the QU4LITY RA. Furthermore, they can benefit from the Digital Models of the project, especially in deployments that require data exchange and interoperability across more than one platform.

2.2 Overview of Requirements from Pilots

The Big Data infrastructures and algorithms developed in this deliverable are primarily implemented to support the QU4LITY pilots. Hence, they are destined to address the main requirements of these pilots regarding data-intensive operations, yet they can also be generalized and used in other ZDM settings as well. In following paragraphs, each of the presented platforms and algorithms are matched to the pilot where they belong. In several cases, the rationale for the use of specific algorithms has been also driven by the early inspection of datasets from the production lines of the pilots. Nevertheless, at a higher level, the results of this deliverable have been also driven by the data-intensive characteristics of the various pilots, which are outlined in deliverable D2.1 and detailed in the trials handbooks that are developed in WP7 of the project. The following table outlines these data-intensive requirements for each one of the pilots.

Pilot no.	Company	Title
I.1	Philips	Creating a competitive factory by application of autonomous quality on a OneBlade shaving unit production line
High-Level Data-Driven Requirements: The pilot involves the analysis of data associated with processes signals and dimensional CTQ (Critical To Quality) components in order to predict the quality of the products. It will also attempt to learn predictive quality indicators based on unknown data. Both predictive processes involve data mining processes that will aim at identifying the sources of the output variations that lead to defects, but also how the source variations relate to the observed output variations. The data driven process involves the extraction of knowledge about which sources cause the defects as a means of placing and deploying proper sensors in the line in order to track them and predict quality accurately.		
I.2	Siemens	Data analytics for ZDM efficiency increase in in electronics production

High-Level Data-Driven Requirements: The pilot involves the analysis of quality-related datasets towards (automatically) classifying identified failures and increasing the transparency of the quality management process. There are labelled data available that can be analysed based on supervised machine learning algorithms, yet unsupervised learning based on non-labelled data can be applied as well.

I.3	Continental	Autonomous Quality in PCB Production for Future Mobility

High-Level Data-Driven Requirements: The pilot involves collection and analysis of quality-related data from multi-stage production lines. The developed system captures, communicates, stores and visualizes real-time data on products, material, equipment, environment, human actions and quality and processes. Various data sources from the production line and the warehouse infrastructure will be integrated. The pilot requirements are not directly addressed by the data mining algorithms presented in this deliverable.

I.4	Whirlpool	Dryer Factory Holistic Quality Platform

High-Level Data-Driven Requirements: The main data driven requirements for the Clothes Dryer production concerning the integration and consolidation of data from diverse data sources, with less requirements for advanced predictive analysis. Hence the pilot is addressed by other data-driven digital enablers of the project (e.g., the consolidation and use of a common and holistic semantic model for quality data representation) rather than by data mining algorithms of this deliverable.

I.5	Mondragon	Zero defect & Autonomous Quality in Machinery Building for Capital Goods sector

High-Level Data-Driven Requirements: One of the pilot scenarios for this pilot involves the implementation of process control on hot stamping equipment. This involves the extraction of knowledge for the reasons which cause defects as a means of optimizing related processes that rely on hot stamping (e.g., boron steel components manufacturing). As part of the same pilot (yet in another scenario), quality data analysis for precision machining (i.e. in Cutting/Grinding Machinery) will be performed, by means of the establishment of a multiple data source DSS (Decision Support System). One of the platforms of the present deliverable is destined to enable data collection from multiple sources in support of this DSS.

I.6	Kolektor	Autonomous detection and removal of defects in injection moulding

High-Level Data-Driven Requirements: This pilot involves (nearly) real-time (predictive) data analytics in order to anticipate and timely remove the cause of the process failures in the moulding process. Data analytics are required for anomaly and failure detection and leverage QU4LITY digital enablers (e.g., fog nodes) beyond the scope of this deliverable.

I.7	Thyssenkrupp	Quality Management of Steering Gear based on Acoustic control

High-Level Data-Driven Requirements: The pilot aims at improving the quality management process based on a better correlation between process and control. From a data viewpoint, there is a need for collecting data from various points on the production line, so as to capture in-process sensor data, data from production monitoring and operation data. Machine learning and data mining techniques are to be employed for understanding the correlations between process and control, including their impact on the product quality. A data mining approach that enables the identification of high-confidence rules for such correlations is developed and presented in this deliverable.

I.8	Airbus	Trade space framework for Autonomous Quality Manufacturing Systems' Design

High-Level Data-Driven Requirements: The pilot involves collection of data from various sources (at high level: from the Supplier network, the Factories, Machines and Processes). The data will be analysed to identify and maintain quality targets at pertinent points.

I.9	GHI	Real-time cognitive hot stamping furnace 4.0

High-Level Data-Driven Requirements: For a data perspective, the pilot involves data-driven monitoring of the industrial furnace operating parameters, as a means of identifying sources and cause of the quality issues and subsequently reducing the defective parts.

I.10	RiaStone	Autonomous Quality ZDM implementation for "Ceramic tableware Single-firing"

High-Level Data-Driven Requirements: The pilot is implementing a data driven DSS for quality management decisions, including automated quality control processes in ceramics (e.g., automated weight control in green ware and automated in line control of glaze rheology properties (density, glaze temperature, viscosity)). Data analytics techniques will be implemented in order to timely raise early detection warning signals of production factors that are sources of product defects. The pilot will leverage data collection and data integration platforms that are presented in this deliverable, along with data mining techniques that will be implemented on top of them.

I.11	PRIMA	Additive Manufacturing Pilot Adaptive Control Technology (AMPACT)

High-Level Data-Driven Requirements: The pilot develops a system for collecting, tracking and analysing data to enhance process robustness in Additive Manufacturing (AM). The system emphasizes modularity and benefits from in-line sensor visualization techniques for large datasets (e.g., materials data), which are developed and detailed in this deliverable.

I.12	Danobat	High precision machining – Danobat cutting/grinding machine-tools

High-Level Data-Driven Requirements: This is another precision machine pilot that emphasizes sensor data collection and analytics on Danobat's cutting/grinding machine-tools. It benefits from prediction of the machine's Remaining Useful Life (RUL) as a means of anticipating and alleviating problems in the operation of the machine. The present deliverable develops and report various RUL calculation methodologies.

I.13	Fagor	Zero-Defects Manufacturing Digital Press Machine

High-Level Data-Driven Requirements: This pilot collects and analyses critical parameters for the operation of the machinery, in order to analyse the in conjunction with the production process rather than examining each parameter alone. The pilot requires the collection and combination of information from multiple sources and will leverage one of the platforms (OpenVA) that are developed and described in this deliverable.

I.14	GF	GF Digital machine and part twins for zero defect manufacturing
		High-Level Data-Driven Requirements: The GF pilot focuses on the detection, diagnosis, and alleviation of operational issues of robotized machining cell, including deviations on accuracy, productivity and sustainability. It leverages a data-driven approach that aggregates and combines information from Milling and Electro erosion machinery health with information about process performance and geometrical parts characterization. Its focus is on the aggregation and integration of information from multi-stage processes in a common data space. Hence, it is less focused on data mining and not supported by the algorithms that are presented in this deliverable.

Table 1: Overview of High-Level Data Driven Requirements of the Fourteen (14) QU4LITY Pilots

Based on the above-listed descriptions, the following remarks can be drawn:

- All the pilots comprise data-driven operations and rely on the collection and analysis of digital data about products, machines, production processes, control processes and more.
- Several of the pilot will leverage Big Data platforms (i.e. middleware platforms) towards facilitating and accelerating their technical development, but also their configuration and deployment. At least five (5) of them are served by the platforms that are described in this deliverable.
- Most of the pilots (i.e. eight (8) pilots) leverage advanced data mining techniques (including machine learning and deep learning) to extract and product knowledge about the quality management processes and to provide a foundation about ZDM. These pilots will be supported by the Big Data analytics algorithms that are presented in later subsections. Nevertheless, there also pilots that focus on other aspects of a digital infrastructure for industrial use cases (e.g., edge computing, semantic interoperability), which will be primarily served by other digital enablers of the QU4LITY project.

Following sections illustrates the Big Data platforms and data mining algorithms that are explored, developed and used by the WP3 QU4LITY partners, while providing information on the pilots that they will be used.

3. QU4LITY Big Data Platforms

3.1 Overview of Big Data Platforms

As part of WP3 of the QU4LITY project, three Big Data platforms will be deployed and use in various pilots. These platforms leverage the state-of-the-art Big Data infrastructures (e.g., Apache Kafka) and IoT capabilities (e.g., IoT Gateways), yet they also provide added-value features for data pre-processing and analytics. In-line with the QU4LITY Reference Architecture, these platforms will serve as hosting environments for the data analytics and machine learning techniques that are described in the following section.

The following table provides an overview of the three BigData analytics infrastructures / platforms that will be used in the project, while following paragraphs elaborate on their capabilities and the ways they are customized and used in the scope of the QU4LITY project.

Platform	Description	IP Owner
FAR-EDGE DDA	Supports Routing and Preprocessing of Heterogeneous Data Streams in Industrial Environments	AIT
Open VA	Supports Analytics and Visualization of IoT Data	VTT
Data Fabric	Supports data capture and ingestion, including real-time streaming, transformation, storage and indexing; Analytics and machine learning	LKS

Table 2: Overview of QU4LITY Big Data Platforms

3.2 FAR-EDGE Distributed Data Analytics (DDA) Infrastructure

3.2.1 Description

3.2.1.1 Background Development

Overview

The FAR-EDGE DDA is a configurable infrastructure for Distributed Data Analytics, which enables the collection, aggregation, integration and preprocessing of diverse data streams. It was developed in the scope of the H2020 FAR-EDGE project and is available as open source through the Edge4Industry (www.edge4industry.eu) community. In practice, it is a distributed middleware infrastructure, which is agnostic of the underlying data sources and streaming analytics middleware toolkits. As such it provides the means for configuring and deploying streaming analytics applications over a variety of heterogeneous data sources, including streaming analytics engines such as Apache Spark and Kafka.

As outlined in Figure 3, the FAR-EDGE DDA is structured in tiers in-line with most reference architectures for IoT applications, including an edge, a cloud and a ledger tier. The edge tier provides the means for low-level, real-time analytics operations that are executed close to the field, typically in the scope of a local area network. The FAR-EDGE DDA comprises an Edge Analytics Engine (EA-Engine), which is instantiated and executed at the edge of the network. On the other hand, the cloud tier enables the execution of analytics that span multiple edge nodes (and EA-Engine instances) by means of a cloud-based Distributed Analytics Engine (DA-Engine). Hence, the DA-Engine executes analytics functions over data streams that are aggregated in the cloud. It can handle larger amounts of data than the EA-Engine yet is not suitable for real-time low overhead stream processing. The FAR-EDGE DDA introduces also a ledger tier, which provides the means for decentralized configuration and synchronization of edge analytics processes, using a distributed database based on blockchain technology.

Edge Analytics Engine (EA-Engine)

The EA-Engine is deployed and executed close to the field i.e. within an edge gateway. Its operation is based on a Data Routing and Pre-processing (DR&P) component, a data bus and a registry of devices (i.e. Device Registry). Specifically, the DR&P component routes data from the data sources (e.g., industrial platforms and devices that may include data streaming platforms) to the Edge Analytics Engine (EA-Engine). The routing operations are based on information contained within the Device Registry, which contains information (e.g., connectivity protocol, IP address and port) about how the various devices and data sources can be accessed. The registry ensures the dynamism of the EA-Engine, data sources can at any time register or deregister from the device registry. Moreover, the component provides pre-processing capabilities, which allow for transformations to data streams prior to their delivery to the EA-Engine.

In addition to interacting with the Device Registry, the DR&P component provides a Data Bus, which is used to route streams from the various devices to appropriate consumers, i.e. processors of the EA-Engine. The Data Bus is not restricted to routing data streams stemming directly from the industrial devices and other shopfloor data sources. Rather it can also support the routing of any data streams and events that are produced by the EA-Engine. Overall, the EA-Engine is a configurable runtime environment hosted in an edge gateway, which executes data analytics close to the field in order to meet stringent latency requirements.

Towards configuring the operation of the EA-Engine, DDA application developers and solution providers can edit a specification of analytics tasks, which is provided as an XML (eXtensible Markup Language) file and serves as a DSL for customizing EA-Engine's operation. This DSL provides the means for expressing streaming analytics functions based on the combination of several processing functions that are conveniently called "processors". The latter processing functions operate over streaming data that are available in the Data Bus of the DDA.

The EA-Engine supports three different types of processors, namely:

- **Pre-processors**, which pre-process (e.g. filtering) data streams and prepare them for analysis by other processors. Pre-processors acquire streaming data through the DR&P component and produce new streams that are made accessible to other processors and applications through the Data Bus of the infrastructure.
- **Storage processors**, which store streams to some repository such as a data bus, a data store or a database. They provide the persistence functions, which is a key element of any data analytics pipeline.
- **Analytics processors**, which execute analytics processing functions over data streams ranging from simple statistical computations (e.g., average or a standard deviation) to more complex machine learning tasks (e.g., execution of a classification function). Like pre-processors, analytics processors can access and persist data to the Data Bus.

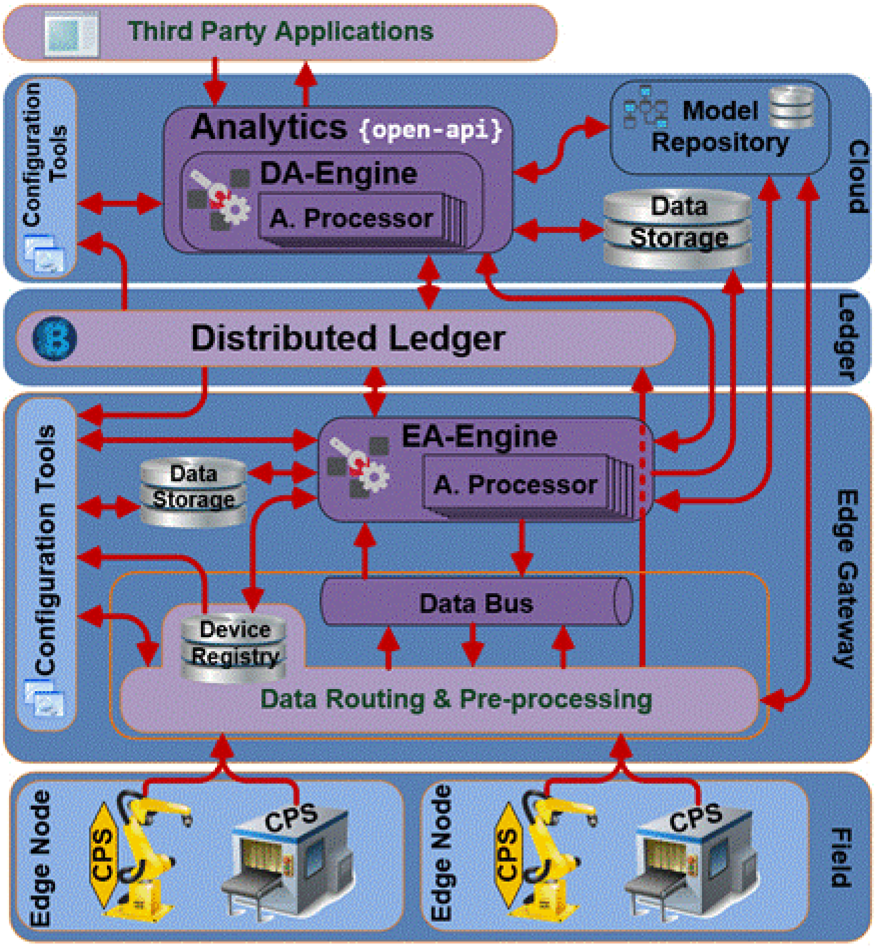


Figure 3: Anatomy of the Distributed Data Analytics Engine

The EA-Engine can execute analytics pipelines comprising combinations of these three processors (as shown in Figure 3). The relevant pipelines (or workflows) are described through well-defined configuration files, called Analytics Manifests (AMs), which essentially represent the important part of the DSL of the EA-Engine. Specifically, an AM defines a set of analytics functionalities as a graph of processing functions that comprises the above three types of processors and which can be executed by the EA-Engine. AM instances are built based on the devices, data sources, edge gateways and analytics processors that are available. All these devices and data sources are model based on the digital modelling approach that is described in the following Section of the paper and which specifies additional constructs of the DSL of the DDA.

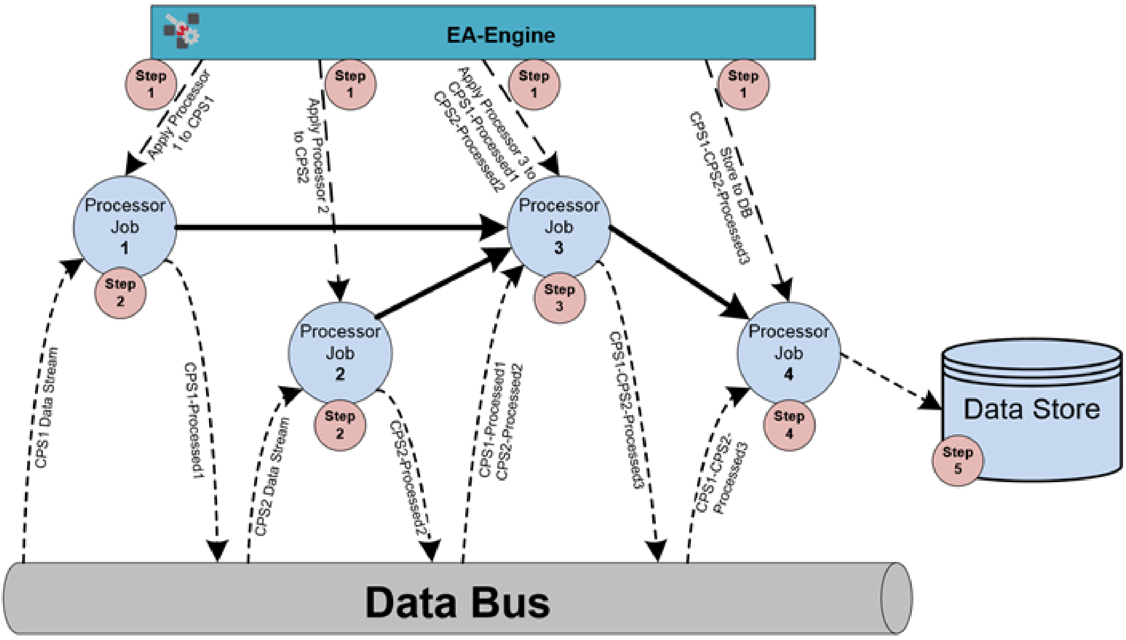


Figure 4: The Different Data Processors of the DDA

An EA-Engine is in charge of specifying and executing edge analytics pipelines comprises combinations of the above-described processing functions. Factory wide analytics comprising multiple edge analytics workflows in a higher-level pipeline are then executed through the distributed analytics engine (DA-Engine) that resides in the cloud layer of a DDA deployment.

Distributed Analytics Engine (DA-Engine)

The DA-Engine is destined to execute global analytics functions based on analytics configurations that span and combine multiple edge analytics instances. It is also configurable and programmable thanks to its support for a DSL that describes global analytics functions in terms of edge nodes, edge gateways, data sources and the processing functions that are applied over them. In particular, the DSL is specified as an Analytics Manifest (AM) for global analytics, which can comprise multiple edge analytics instances specified as AMs as well. Similar to the EA-Engine the DA-Engine leverages the descriptions of all these artifacts within a digital models' repository, which comprises the digital representation of the devices, data sources and edge gateways that are part of the DDA. The structure of these Digital Models is also described in the following section. Note that all the digital models are kept up to date and synchronized with the status of the DDA's elements. Hence, they are accessible from the DR&P, the EA-Engine and the DA-Engine components and can be used in the specification of both edge analytics and global analytics tasks. As already outlined, the DA-Engine stores data within a cloud-based data storage repository, which persists the results of global analytics tasks.

Open API for Analytics

The FAR-EDGE DDA infrastructure defines, implements and exposes an Open API (Application Programming Interface). This API enables external systems to access and integrate the functionalities of the DDA infrastructure, including the configuration, execution and deployment of factory-wide analytics tasks, that span multiple edge gateways. The Open API enables solution integrators to configure the DDA and to execute data processing and analytics functions over data streams stemming from all devices that registered in the registries of the DR&P components of the edge nodes of the DDA. In this way the DDA infrastructure can be used by third-party applications. In practice, the Open API supports the following three types of operations:

- CRUD (Create Update and Delete) operations for AMs, which enable the management of AMs.
- Access to and management of information about deployed instances of analytics workflows, including access to information about AMs and their status.
- Management of a specified AM, including starting, stopping, posing and resuming the execution of an analytics instance expressed in the AM DSL.

The Open APIs has been specified and implemented as a RESTful API, which facilitates their use by application developers and solution integrators.

FAR-EDGE Digital Models Solution

FAR-EDGE has defined a digital modelling solution (as a set of XML schemas) that can represent Distributed Data Analytics (DDA) applications in a way that addresses the interoperability, data routing and dynamic discovery challenges. It also enables the definition of DDA applications based on different sources, thus acting as Domain Specific Language (DSL) for DDA in edge/cloud environment. The solution comprises data schemas for the following (main) entities:

- DSD (Data Source Definition): Defines the structure of a data source.
- DI (Data Interface specification): Defines the interface for accessing the data source.
- DK (Data Kind): Defines the semantics of the data source.
- DSM (Data Source Manifest): Describes an instance of a data source that is represented through a DSD.
- APD (Analytics Processor Definition): Defines analytics functions over data sources.
- APM (Analytics Processor Manifest): Describes an instance of an analytics functions that is represented through APD.
- AM (Analytics orchestrator Manifest): Provides the DSL for defining DDA workflows.

The complete specification of the above-listed schemas is part of deliverable D5.5 of the H2020 FAR-EDGE project.

FAR-EDGE has also provided a compete run-time environment (i.e. a DDA infrastructure) that can process data sources and analytics workflows that are defined based on the above-listed schemas. The DDA infrastructure is available at: https://github.com/far-edge/distributed-data-analytics

The main benefit for using the FAR-EDGE Digital Models and the DDA infrastructure is the interoperability and discoverability of the involved data sources. Using the Data Models and the DDA infrastructure one can represent Logical configurations of edge/cloud systems (e.g., Edge Gateways, Data Sources, Devices), along with their configuration (e.g., definition of new data sources, association of data sources to edge gateways) using IT APIs and tools. Additionally, the DDA infrastructure provides the means for introducing third party analytics processors in a way that renders them reusable and manageable. This ensures the dynamic configurability and extensibility of the platform.

3.2.1.2 Foreground Development

The FAR-EDGE DDA will be customized and used in two pilots of QU4LITY, namely the THYS and RIASTONE pilots. For each customization of the DDA, the following steps will be undertaken:

- **Identification and Modelling of Data Sources**: As part of this step, the data sources will be appropriately modelled based on the specification of proper DSD, DK and other digital modelling constructs.
- **Connecting to Data Sources:** This step involves the specification of DI constructs and the implementation of connectors to the data sources of each pilot.
- **Specification and Implementation of Data Analytics Pipelines**: This step involves the specification and implementation of end to end analytics pipelines, based on appropriate Pre-Processors, Storage Processors and Analytics Processors. This step connects the platform to ML/DL algorithms for ZDM that are presented in the following section (e.g., QARMA or RUL Calculation algorithms).
- **Specification and Implementation of the Analytics Manifests (AM) for the analytics tasks at hand**: As part of this step, the AM for the analytics tasks of each use case will be created and deployed based on the Open APIs of the DDA.
- **Customizing of the Dashboard to the needs of each pilot**: This step involves the creation of a customized dashboard for each use case, based on the customization of the standard/default dashboard of the DDA.
- **Controlling the DDA Instance**: The DDA instances will be managed using the Open API for analytics.

3.2.2 Alignment to QU4LITY RA and Status of Prototype Implementation

The FAR-EDGE platform belongs to the Digital Infrastructures of the QUALITY RA. It provides a solution for BigData and the IoT hub building blocks of the RA. Note that FAR-EDGE provides also the means for implementing data-driven digital automation processes as a means of supporting other functional building blocks of the RA, such as the Digital Twins and Learning Modules of the RA at the Factory level. However, the deployment and use of the platform in the scope of the project will not take advantage of existing learning models and digital twins. Rather, new models will be built based on the requirements of the QU4LITY pilots (i.e. THYS, RIA) that will take advantage of the functionalities of the FAR-EDGE DDA. As already outlined, FAR-EDGE provides a set of digital models for modelling and managing analytics functionalities, which do not however comprise digital twins for ZDM applications.

The customization process of the DDA for the two pilots is at its early stages at the time of writing of this document. As a first step, sample datasets from the pilot sites are processed in order to identify suitable data mining techniques for deriving insights like RUL information. The algorithms to be identified will drive the implementation of the analytics pipelines based on the DDA infrastructure.

3.2.3 Use in Pilots

The FAR-EDGE DDA platform will be customized and used in the scope of the Thyssenkrup and Riastone pilots of the QU4LITY project.

3.3 OpenVA

3.3.1 Description

3.3.1.1 Background Development

VTT OpenVA platform can be used as building block of different integrated analytics applications. OpenVA is a platform that has been used in several EU funded and national research projects to integrate different data sources into a cloud-based analytics application.

VTT OpenVA consist of open source software components that forms a platform of:

- A database that stores the application metadata and measured data in a standard domain independent format. Measured data can be also be stored in external data storage and accessed for analysing with help of the metadata.
- An extendable analysis and visualization library containing a selection of AI analysis and visualization methods. The library is customized based on QU4LITY pilot needs.
- Embedded R and Python statistical computing environments.
- A web user interface where the user can select variables for analysis and explore the data with the help of visualizations. The visualizations can be in e.g. 2D, 3D and interconnected with real object visualizations, such as CAD models. The user interface suggests user the appropriate analysis methods letting them to concentrate on the substance instead of data analysis methods.

3.3.1.2 Foreground Development

VTT OpenVA platform is a fully functional platform for analyzing different types of data and it has been published in open source. Before QU4LITY OpenVA has been used in several business areas, but manufacturing data analysis has been implemented in only small demonstrations.

In QU4LITY, OpenVA is used in a highly configurable manufacturing system environment. That requires implementation of configuration user interface, which will replace current method where configuration is made using SQL commands to change metadata stored in the OpenVA relational database.

In QU4LITY, we will also expand existing analysis algorithms in VTT OpenVA platform by implementing new analysis algorithms that are specific to the QU4LITY domain.

3.3.2 Architecture

VTT OpenVA is based on three-layer architecture with well-defined programming interfaces between the components (Figure 2).

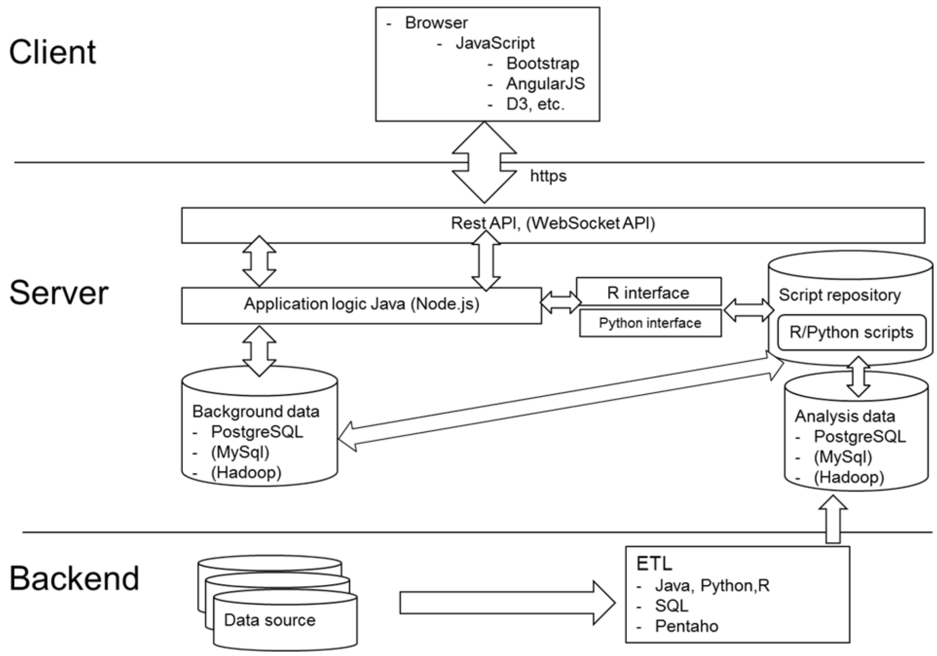


Figure 5: Architecture of the OpenVA Platform

3.3.3 Alignment to QU4LITY RA and Status of Prototype Implementation

VTT OpenVA provides a platform for building pilot analytics services. It aligns to the digital infrastructures concept of the QU4LITY RA, with emphasis on the BigData Analytics part of the digital infrastructures. At the time of writing this deliverable, configuration user interface requirements for OpenVA have been specified. The implementation of the platform and its used in manufacturing processes has started.

3.3.4 Use in Pilots

OpenVA will be in the FAGOR Pilot focusing on manufacturing Processes with Hot Stamping Machines.

3.4 Data Fabric / Big Data Infrastructure

3.4.1 Description

Datafabric is an Industry 4.0 platform for SME industrial companies with capabilities of data capture and ingestion: real-time streaming, transformation, storage and indexing; Analytics and machine learning.

3.4.1.1 Background Development

Datafabric development started in 2018 and during these years have progressively added functionality. Before starting QU4LITY project, LKS developed main components of Datafabric, the control panel and the main platform. Also, the Datafabric Internet of Things (IoT) Gateway was designed and the development was started.

3.4.1.2 Foreground Development

During the project, the development of Datafabric IoT Gateway will be completed and additional functionalities may be added to the pilots if they are detected. On the other hand, a custom Identity and Access Management software piece based on KeyCloak is being developed. This component is called Datafabric Authenticator.

3.4.2 Architecture

The conceptual map of the S.I below shows us a high-level view of the system, where the different layers that make it up are displayed.

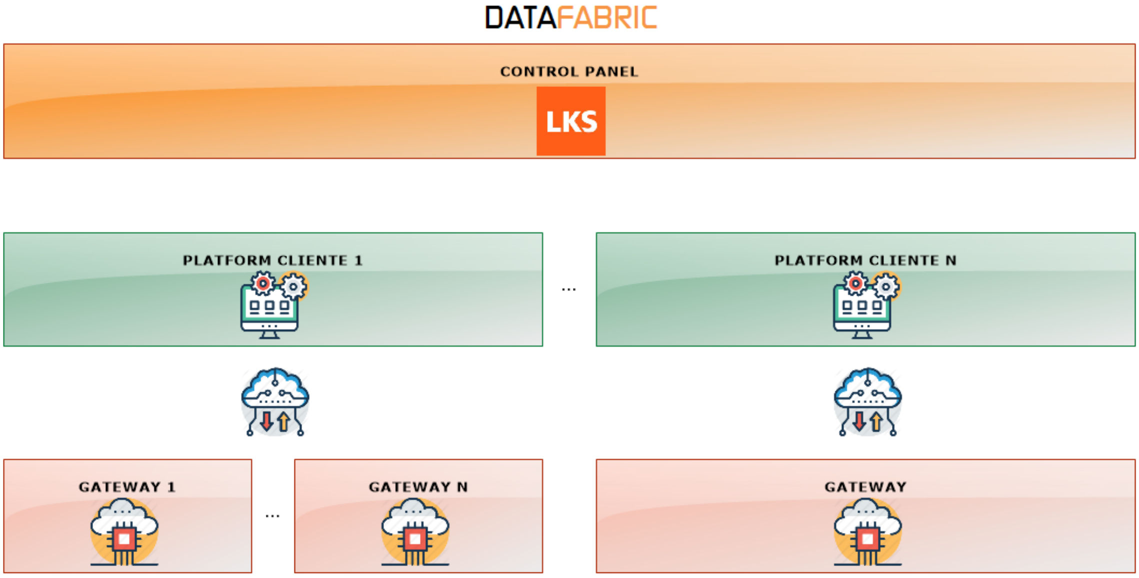


Figure 6: Overview of DataFabric Infrastructure

Below is briefly indicated the mission/responsibility of each layer of the system.

3.4.2.1 DATAFABRIC CONTROL PANEL

This layer will provide LKS with the control capabilities of the entire Datafabric platform.

3.4.2.2 DATAFABRIC PLATFORM

Datafabric Platform comprises both the configuration of the system (structure, devices, assets, sensors, events...), as well as the ingestion of the information captured from the Datafabric Gateways for further exploitation by different services:

- Datafabric Historic is the service responsible for the persistence of information.
- Datafabric Portal are the set of tools that will allow you to monitor events in real time and define alarms, available for computers, mobile devices and flat screens that can be placed in different areas of the manufacturing line.
- Datafabric Dashboard is a business intelligence tool to analyze historical information and facilitate decision making.
- Datafabric Predictive allow you to use different ML models (linear regressions, decision trees, etc.) previously trained, which will contain input variables (causes) and one or more properties of the part (effects), being able to define new events that can be added to the dashboard to define indicators, alarms, etc. The construction of the different models will be specific to each client and dataset.
- Datafabric Authenticator is a custom Identity and Access Management software piece based on KeyCloak.

3.4.2.3 DATAFABRIC GATEWAY

The Datafabric Gateway is an on-site hardware appliance that will allow you to:

- Capture sensor data from diverse data sources.
- Preprocess data (Edge computing).
- Send the information to the linked Datafabric Platform.

3.4.3 Alignment to QU4LITY RA

LKS Datafabric provides platform for building pilot authentication services in FAGOR Pilot with a component called Datafabric Authenticator. Also, the IoT gateway could be used in MONDRAGON Process pilot.

3.4.4 Status of Prototype Implementation – Use in Pilots

The status of the prototype implementation is according to the plan. Development is being executed and integration in the pilots is pending to the evolution of them. The platform is deployed and customized for use in the scope of the Mondragon pilots.

4. QU4LITY Library of Analytics Algorithms

4.1 Overview of Analytics Algorithms

The following table provides an overview of the QU4LITY algorithms for data-driven extraction of ZDM insights.

Algorithms	Description	IP Owner
Data Driven RUL Calculation	Long Short-Term Memory (LSTM) approach for RUL estimation, which can make full use of the sensor sequence information and expose hidden patterns within sensor data with multiple operating conditions, fault and degradation models.	ATLAS
Model Driven RUL Calculation	Performs distribution fitting to the available data and can be combined with data driven RUL	ATLAS
Deep Learning for RUL Calculation	Use of LSTM and Attention-based Networks	AIT
Quantitative Association Rule Mining (QARMA)	Data Mining approach that produces quantifiable rules based on the sets of features that appear frequently together in the training dataset	AIT
Decision Support for Quality Sensors' Installation	Determines where measurements should be placed in a production line, in order to identify/verify a data-driven input-output model	TNO
Analytics for In-Line Sensor Data Visualization – Design4AM	Techniques for Intuitive volumetric modelling, mechanical simulation and interactive visualization of organically shaped parts	FHG-IGD

Table 3: Overview of QU4LITY Analytics Algorithms for ZDM

Following paragraphs provide details about each one of the above algorithms.

4.2 Hybrid Remaining Useful Life Calculation

4.2.1 Overview

The Remaining Useful Life (RUL) analysis can provide valuable information regarding the deterioration rate of assets as the former is defined as the length from the current time to the end of the useful life. Accurate RUL estimation plays a critical role in the improvement of the quality of the produced product and in the Zero-Defect Manufacturing process in general. The RUL analysis can be either model-driven or data-driven. In the context of the Qu4lity project, ATLANTIS Engineering will deploy solutions from both the algorithmic families.

4.2.1.1 Model-driven RUL calculation

The RUL model inputs are condition indicators, features extracted from sensor data or log data whose behavior changes in a predictable way as the system degrades or operates in different modes. Depending on the provided data the RUL is calculated in different forms.

In the early stages of data collection, where there are not enough historical data collected and no or not enough run-to-fail data, the appropriate RUL estimation form is the Degradation (or Deterioration) models [Lu97] using threshold information. The second form of RUL estimation uses lifetime data [Si11], which are sensor data from the lifetime of the equipment (from the normal functionality until the end of life). The last form of RUL estimation utilizes run-to-failure histories of machines similar to the one which is examined [Peng09].

A crucial phase of the RUL estimation is the distribution fitting to the available data, as the estimation results might considerably differ between various selected distributions. There are several approaches, which provide statistical information considering the goodness of fit [D'Agostino86] of a distribution to the specified data. Most known techniques include the Kolmogorov-Smirnov (K-S) [Chakravarti67], Anderson-Darling [Stephens74] (modification of the K-S test), Chi-Square [Snedecor89] , Akaike information criterion [Sakamoto86], Maximum Likelihood Estimation (MLE) [Myung03], Coefficient of Determination (R2) [Barrett74], Posterior Predictive P-values (PPPs) [Gelman03] and Sum of Square Error (SSE) [Motulsky87]. Although for some tests there are known limitations (e.g. the K-S is appropriate only for small scale of data), none of the above dominates the others. In the QU4LITY projects context ATLANTIS Engineering will provide a benchmark solution to assist the distribution selection phase, which will rank different known distributions based on their fitting scores in more than one tests (proving a configurable weighting scheme between the available goodness of fit tests).

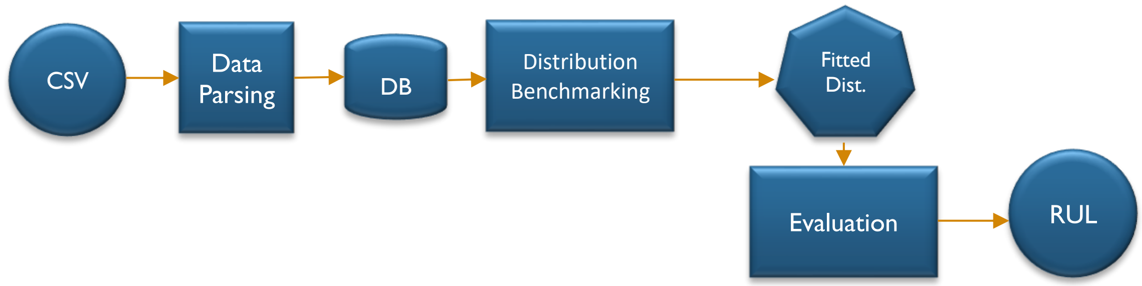


Figure 7: Model-driven RUL Calculation Process.

Figure 7 presents the process of the of the model-driven RUL calculation. The sensorial data are provided in .csv format and they are parsed into a timeseries database (DB). Through the proposed distribution fitting benchmarking approach (Distribution Benchmarking), the distribution with the highest rank is selected as the most appropriate one (Fitted Dist.). Based on the current state and the selected distribution the RUL is calculated.

4.2.1.2 Data-driven RUL calculation

Data-driven approaches for RUL estimation, like regression-based or Convolutional Neural Network (CNN) approaches use features created from sliding windows to build models. However, sequence information is not fully considered in these approaches. Sequence learning models such as Hidden Markov Models (HMMs) and Recurrent Neural Networks (RNNs) have flaws when modeling sequence information. HMMs are limited to discrete hidden states and are known to have issues when modeling long-term dependencies in the data. RNNs also have issues with long-term dependencies.

ATLANTIS Engineering will use the Long Short-Term Memory (LSTM) approach for RUL estimation, which can make full use of the sensor sequence information and expose hidden patterns within sensor data with multiple operating conditions, fault and degradation models.

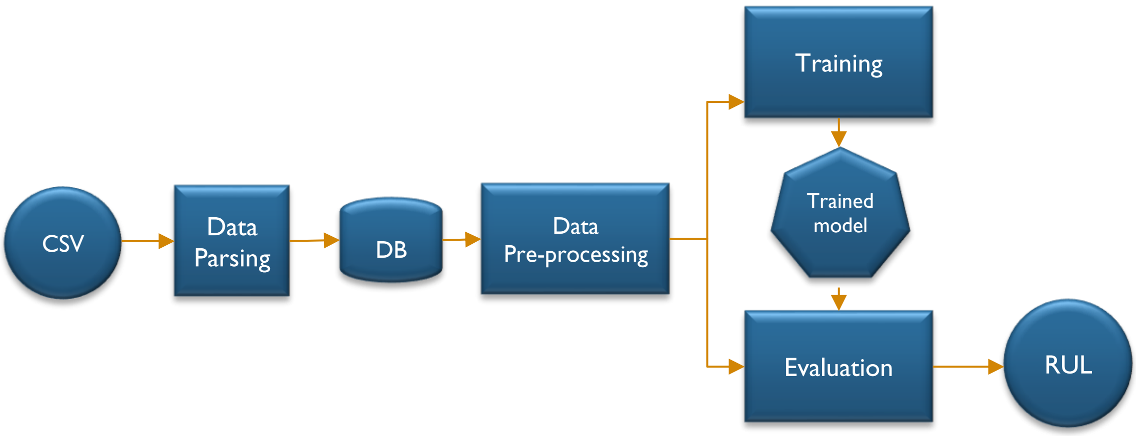


Figure 8: Data-driven RUL Calculation Process.

As Figure 8 presents, the process is similar to the model-driven approach up to the point of storing to the timeseries database. In order to train an LSTM model for RUL calculation, special data pre-processing should be applied to the stored data. The result of the pre-processing is a training set appropriately formed to map the evolution of the RUL as the system approaches the end of life. The training dataset is fed to the LSTM neural network and a trained model is produced. The model is used to calculate the RUL from the current state of the system.

4.2.2 Use Cases – Examples

The presented algorithms are destined to be applied in the Danobat Railway Systems and more specifically, the grinding machine presented in Figure 9.

Figure 9: Danobat Railway Systems - Grinding machine.

The main components of the grinding machine for the use case, are presented in Figure 10. A workpiece is taking form by a grinding wheel supported by a regulating wheel. The two wheels are supported by dresser wheels which help to expose fresh grains and to maintain their profile.

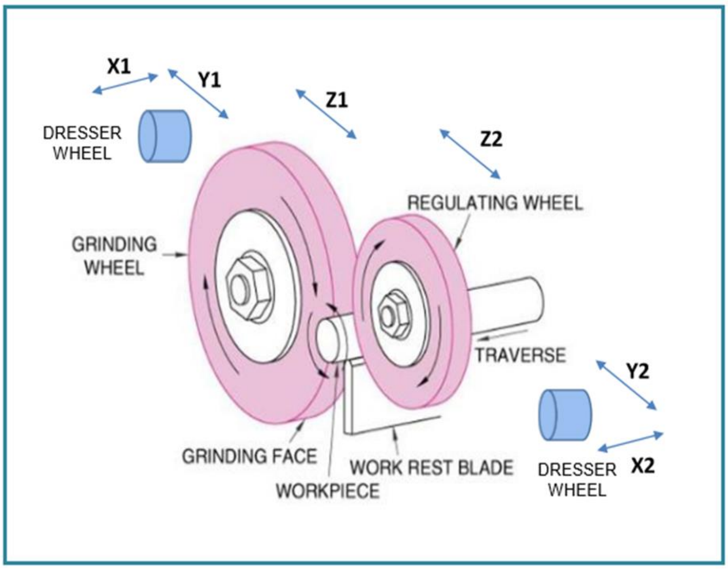


Figure 10: Danobat Railway Systems - Grinding parts.

The obtained data are measurements provided by sensors installed on the machine. Multiple types of measurements are provided, like the Intensity of the wheels' axes, shown in Figure 10, their temperature, the load of the machine and other measurements related to the general functionality of the machine. Added to the sensorial measurements, the dates of failure incidents are provided with focus on the failures reported on Z2 axis.

The objective is to provide RUL calculations either in working time manner or in processed parts manner. More details on the RUL calculations will be provided as part of the detailed trials descriptions in WP7.

4.2.3 Customization for QU4LITY

The model-driven RUL analysis is not the most appropriate approach for handling Big Data due to limitations in the distribution fitting process. Hence, this approach can be used to assist the initial analysis of specific cases, where the provided data is not enough for data-driven modeling (i.e. efficient LSTM models training). The results from the model-driven RUL can be utilized as ground truth to evaluate the accuracy of the data-driven approaches.

The RUL calculation based on processed parts needs the added information in the training set of how many parts are processed by the machine. However, this information is not provided. In order to mine this information, the State measurement of the machine and the information that every part is processed in 30 seconds, will be utilized. The State indicates when the feeding of workpieces in the machine is halted (a workpiece is in processing state) or not.

4.2.4 Status of Prototype Implementation

Currently, in the prototype implementation of the model-driven RUL, the SciPy[1] python library is used to test the fitting of multiple distribution using the Sci-Square metric. The benchmark solution for the mixed ranking of the fitting of each distribution will be provided in a future version.

Considering the data-driven RUL approach, the LSTM algorithm has been tested in a preliminary dataset from the Danobat use case. The development roadmap involves more extensive testing with broader datasets and algorithm parametrizations.

4.2.5 Use in QU4LITY

The results from the model-driven approach will be utilized in the training and evaluation phase of the data-driven approach, providing a ground truth dataset. A RUL calculation tool will be implemented, which will use the trained LSTM models for online evaluation.

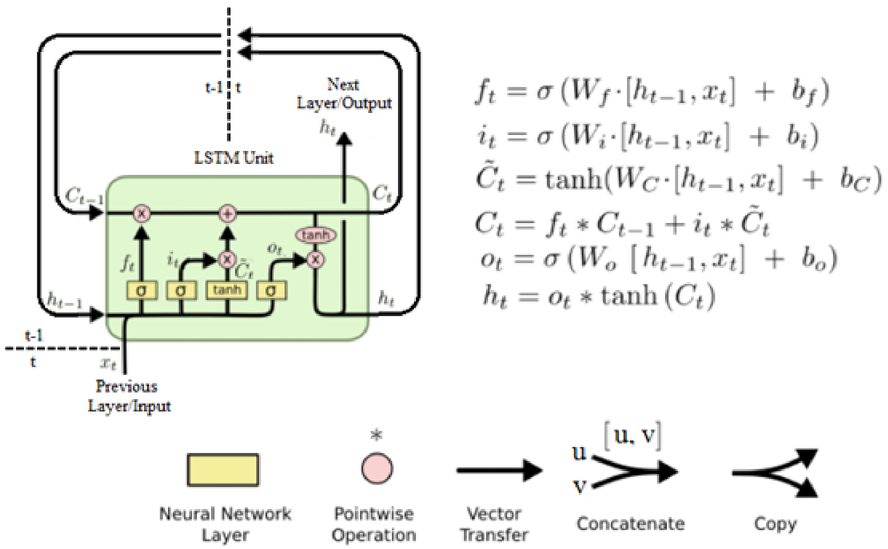
[1] https://www.scipy.org/

4.3 Deep Learning for Remaining Useful Life Calculation

4.3.1 Overview

Recurrent Neural Networks (RNNs) were introduced in the late 90's to solve exactly prediction problems, as typically use their internal state (memory) to process sequences of inputs. Presently, there are two well-established variants that have proved very robust and efficient in a wide variety of tasks:

- **Long Short-Term Memory Networks (LSTM)** [Sherstinksi18]: they are a variant of classical RNN with so-called forget gates that allow for some "forgetting" of old or non-important data during training. Architecture shown in Figure 11
- **Attention-based Networks** [Luong17]: another more recent variant of RNN that avoid to a greater degree the so-called "curse of dimensionality" of long sequences. They have been particularly successful in sentiment analysis in social networks but also in protein structure analysis tasks. Architecture shown in Figure 12.



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$
$$o_t = \sigma\left(W_o \, [h_{t-1}, x_t] + b_o\right)$$
$$h_t = o_t * \tanh(C_t)$$

Figure 11: LSTM Recurrent Neural Network Architecture Allows for Sequential Input Data

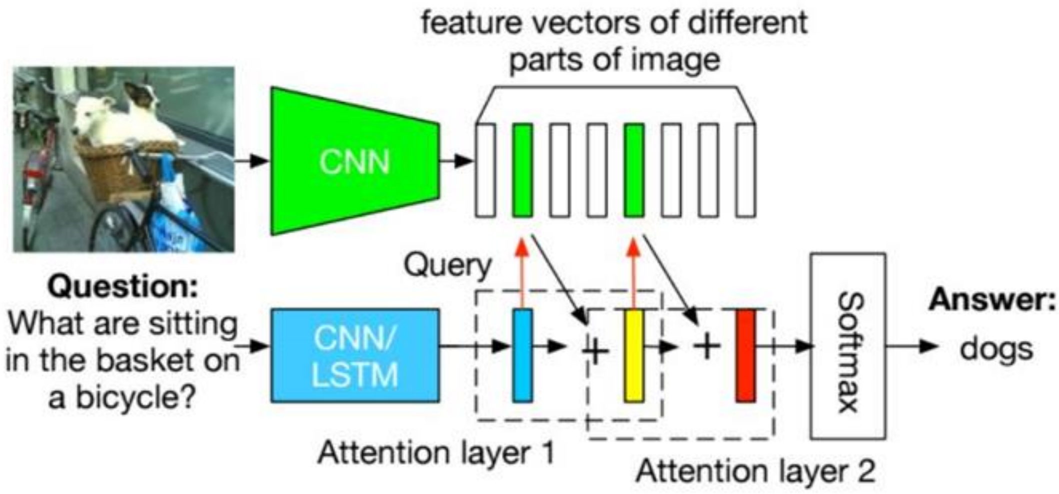


Figure 12: Attention-based Neural Network Architecture

4.3.2 Use Cases - Examples

Models such as the above presented one allow for varying length of input sequence data which is a requirement in the datasets of the Thyssenkrupp pilot. The initial data provided by Thyssenkrupp comprise product measurements and do not include (quality-related) sensor readings. This makes it challenging to predict RUL, given that the latter depends highly on the status of the machine and hence on sensor readings. On the other hand, product-related measurements provide only indirect indicators about the status of the machine and render RUP prediction very difficult. Note however, that in a second round of data collection/provision Thyssenkrupp will also provide sensor data as it is currently installing sensors in the machines. Sensor readings will facilitate the use of RUL algorithms.

4.3.3 Customization for QU4LITY - Status of Prototype Implementation

The presented RNNs variant algorithms have been currently implemented and tested in Thyssencrup datasets without very promising results. Nevertheless, significant improvements are expected once the sensor data will become available. The presented RUL algorithms will be applied also on Riastone datasets.

4.4 Quantitative Association Rule Mining (QARMA)

4.4.1 Overview

QARMA and its variants (e.g. the R4RE system [3, 4] recently published for detecting rare events in Predictive Maintenance settings) represent a drastically different approach to identifying and predicting when the next defect will happen in the scope of sequential datasets (i.e. parts made in a time-sequence). Instead of training a (possibly deep) network of hidden layers of neurons that activate via some kind of step-wise activation function so that eventually the network learns to predict the right target values, QARMA implements a Data Mining inspired approach in which sets of features that appear frequently in the dataset together are collected together, and then each feature is quantified (its value restricted in a numerical interval) with the goal to derive conditions that imply that a target variable among the features takes on a desired value. A relevant Graphical User Interface (GUI) is illustrated in Figure 13.

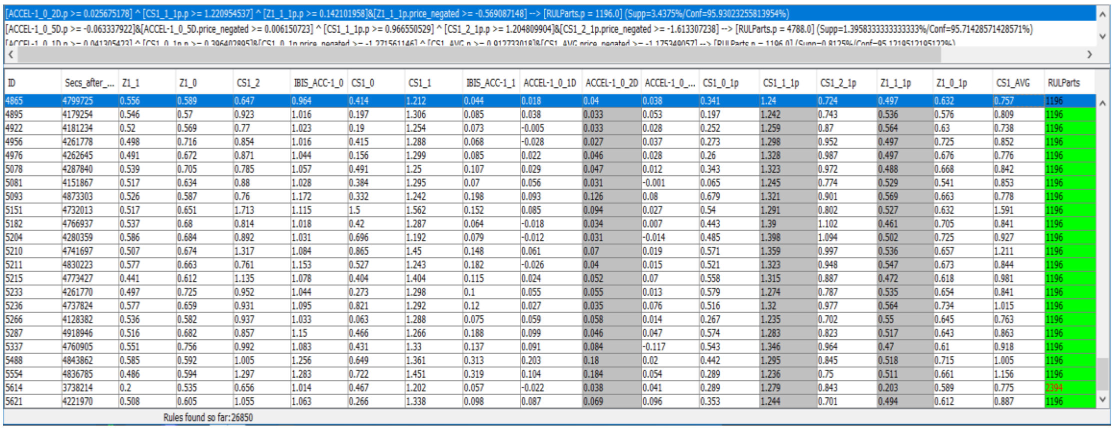


Figure 13: QARMA GUI resembles a spreadsheet with appropriate filters applied, so that one column (the green one) satisfies a certain condition

4.4.2 Use Cases - Examples

The QARMA algorithms have been validated and used in predictive maintenance use cases. As part of the H2020 PROPHESY project[2], AIT has tested the approach with data from the plant of Jaguar Land-Rover (JLR) and has been shown to produce excellent error rates in regression prediction tests. It is also integrated within the PROPHESY Integrated Machine Learning (ML) Toolkit and thus has proved interoperability support with the PROPHESY project IoT infrastructure and analytics ecosystem.

4.4.3 Customization for QU4LITY - Use in QU4LITY

In QU4LITY the QARMA algorithms will be tested and validated in the scope of the Thyssenkrupp and Riastone pilot deployments. In particular, the data provided by Thyssenkrupp are mainly relating to product measurements rather than sensor readings from the machines themselves that produce the products. This characteristic presents a challenge in terms of Remaining Useful Life (RUL) prediction because usually it is the state of the machine that determines when there will be a maintenance need for this machine. Measurements on product characteristics are only indirect indicators of the state of the machine, and thus present extra challenges

[2] https://prophesy.eu/

to the Predictive Maintenance task. Furthermore, the product characteristics measurements are given in the form of a number of curves on the (X,Y) plane for different X- and Y- quantities (e.g. X-axis could be representing degrees, and Y-axis could be representing velocity measured in m/s etc.) To further complicate things, these curves are often of different lengths. The same part undergoes transformation processes in several modules, and each process may result in an OK or NOK (NOT OK) part. In this context, the implementation of the Thyssenkrupp pilot will involve training a system to be able to tell how many parts each module will be able to process before the next NOK part appears in the same module (machine).

4.4.4 Status of Prototype Implementation

As of the time of writing of this deliverable, the QARMA algorithms have been tested in a preliminary dataset from the Thyssenkrupp. The pilot development roadmap involves more extensive testing with broader datasets, as well as the integration of the training algorithms in the pilot system over the FAR-EDGE Distributed Data Analytics platform that will be used for data collection and routing in the scope of the pilot.

4.5 Product Quality Classification

4.5.1 Overview

Machine Learning methods can be subdivided according to the basic principle of their functionality. According to Lieber [2018], these principles are:

- Supervised Machine Learning (e.g. Decision Trees, Neural Networks).
- Unsupervised Machine Learning.
- Semi-supervised Machine Learning.
- Reinforcement Machine Learning.

Supervised and unsupervised Machine Learning processes form the basic framework for predictive and explorative data analysis [Lieber 2018]. With the use of supervised Machine Learning, models are developed that enable the prediction of future expressions of a defined output variable (label). The prediction is based on algorithms that capture the functional relationships in the underlying training data sets. Based on the availability of data where the correct expression of the label is known a priori, it is possible to develop a model with the highest possible generalization capability, which can classify previously unknown instances [Russel and Norvig 2004].

Unsupervised machine learning methods, on the other hand, use unlabeled historical data, so that there is no prior knowledge of specific output values. These learning methods focus on identifying critical patterns regarding structures or correlations (e.g. clusters or outliers) in the underlying database [Russel and Norvig 2004]. Therefore, only the analysis of the partially hidden data structure is of interest.

4.5.2 Use Cases - Examples

During several research and industrial projects (e.g. CSC - Cyber System Connector, Clean, Pro Mondi) the suitability of Machine Learning and Data Mining to increase the overall quality rate of products as well as the testing efficiency throughout the value stream has been proven. These activities base on reliable prediction of the product quality.

4.5.3 Customization for QU4LITY – Use in QU4LITY

In QU4LITY the Machine Learning algorithms will be used in the Siemens pilot to reduce pseudo errors. Pseudo errors occur since the automated quality tests are programmed to control univariate product measurements. However, multivariate correlations may result in defective products that are not yet known. Therefore, many defective declared parts arise which must be verified in a manual visual inspection. In this context the manual visual inspection defines good or defect products which are used as labels for training and validating a Machine Learning algorithm. Since these labels are based on subjective human decisions, they are not objective.

Since these errors make a significant percentage of the overall test results a methodology will be developed which uses Machine Learning algorithms to identify multivariate patterns within automated testing equipment to reduce pseudo errors during testing.

4.5.4 Status of Prototype Implementation

The provided methodology has been tested with an initial historical dataset from Siemens. To build up the efficiency of the methodology a broader dataset will be used for training and validation. Furthermore, the implementation in the pilot system of Siemens will be considered.

4.6 Decision Support for Quality Sensors Installation

4.6.1 Overview

In this section, we will introduce the problem setting and the envisioned scope of the contribution to the QU4LITY project.

Zero defect manufacturing (ZDM) for a production line requires the reduction of output variations to a level where it will always pass quality tests. This is a challenge that is extremely difficult and costly when performance requirements are tight, as these variations are introduced at every step in the production line. Dealing with such variations requires us to answer two questions:

I. What are the sources of the observed output variations that lead to defects?
II. What is the relation between these source variations and the observed output variations?

The questions I and II highlight the chicken-and-egg problem that is inherent to the ZDM challenge. If we can answer (and measure) I, then we can identify II. Or vice-versa; if we know II, then we can identify I.

The problem is not new and one approach to deal with it is through process control. Process control typically considers the challenge of getting a process to always produce the desired output in spite variations that are imposed. The approach is usually executed in several steps (that can be iterated):

1. **System design for controllability**; where the goal is to design a system with simple dynamics that measures significant sources of variations.
2. **Modeling for control**; characterization the so-called feedforward behavior of the system. It facilitates the capability to predict system outputs given the inputs / disturbances. Such a model is typically dynamic.
3. **Control law design**; or the so-called feedback loop. The control-law exploits knowledge on (a measured) output or other parameter to change the process/input so that the desired output is obtained. This is done based on the knowledge on the system behavior expressed in the model from step 2.

The three steps are all highly relevant for zero defect manufacturing and can be considered at different system levels. The implementation of feedback control is in fact known as the de-facto method to deal with the variations in a production process that lead to defects. Answering questions I & II is accordingly done typically done at step 1 and 2 of the process control approach described above. A process model can be obtained by considering fundamental physical laws (related to mechanics, fluid dynamics, thermodynamics and electronics) that are relevant to the system or by using a data-driven approach, where the measured data on inputs and outputs is used to identify the relation. It is furthermore possible to use a combination of the two, where a structure is obtained from the fundamental physical laws and where model parameters and disturbances are identified from data. We can accordingly tackle the problem in two ways:

- Use fundamental physical laws to determine the relations (answering II) and measure the identified variables of interest (answering I).
- Use (expert) system knowledge to determine where measurements should be placed (answering I) and use the obtained data to identify the relation (answering II).

There is merit in both approaches, and they are accordingly both widely used. The former approach is the more traditional one; it requires low computational power and assumes that relevant data not abundantly available. On the other hand, a disadvantage is that it is not clear what level of detail should be incorporated in the model to make predictions that are sufficiently accurate. Making a detailed model in this manner is furthermore very time-consuming. The latter approach puts more demands on the sensors and computing systems and requires, often complex, data-processing steps. A major advantage is that the obtained models are often valid due to the large amount of data used to obtain them. As sensors prices become lower, computing power increases and accuracy tolerances tighten, the latter approach receives increased attention. The hype on machine learning and artificial intelligence is centered around the expectations that the latter approach can provide highly accurate models by identifying non-obvious relations through a generic approach that can be re-applied to different use-cases in a straightforward manner.

4.6.2 Classification of variations

There are several ways to classify variations. We will now discuss classification based on the source of the variations and their relevant dimension.

We typically consider three sources of variations to a production process:

1. Process variations.
2. Input variations.
3. External disturbances.

The distinction is made by considering where a variation enters the system. A process variation: (i) is created during processing, these variations would accordingly cause two identical process inputs to your two different process outputs. An input variation (ii) is exists before the materials are processed. These can be the output variations of a previous station or the characteristics of new materials that enter the production line. Lastly, we have the external disturbances (iii). These variation sources are not related to the products or processes directly, but instead to changes in the production environment. This can for example be caused by differences in humidity or temperatures. The sources of variation are in practice difficult to distinguish. This particularly true for the discriminations between (iii) on the one hand, and (i) or (ii) on the other hand.

Another way to consider variations is by on the dimension in which they can be evaluated. We can generally consider two dimensions:

a) **Time-based**

b) **Event-based**

Variations that exist in the one dimension may not exist in another dimension at all or they can make no sense whatsoever. For example; the painting of a product can be considered as an event, after which this product is placed in a buffer for a time that varies from product to product. Products are painted red and blue in repeated sequence. Trying to understand the spread of the colors of the product based on the time they spend in a buffer will not yield any insight, while it does if you consider the event-based product number. On the other hand, trying to understand how dry the paint is per product number will not give a clear correlation, while the time in the buffer will.

The discriminations between (i), (ii) & (iii) on the one hand and (a) & (b) on the other hand are highly relevant. Attribution of the variations to the wrong category can work counterproductive in the reduction of its influence. Another aspect to consider is the choice of models that can describe the propagation of these variations through a system. Any model used to express this should be able to describe the variations to such an extent that their influence is understood sufficiently for elimination of the defects, if ZDM is the purpose. This is accordingly dependent on product tolerances but potentially poses a very stringent requirement.

Scope of the contribution

The envisioned contribution lies in a methodology use existing knowledge from the human operators on a production line to select measurements to be installed in order to achieve ZDM. This production line scale is accordingly the first limiting scope that we impose. We consider a line that consists of multiple stations, which can be sequential or parallel and can merge or split. Such a production line can be assumed to produce high volume, which in turn emphasizes the importance of the joint statistics of products that move through the line. The aggregation level is accordingly higher than if a single processing station was considered, in which case attention is typically given to changes that occur in a single product during processing.

We expect that most of the relevant joint statistics in the production line will be related to events or a combination of events and time. Such events can be processing, maintenance, a shift change, cartridge changes and so on. For the combination of events and time, one can think of the time passed since last maintenance, for example. Applying this scope requires us to consider a model or system analogy that allows this evaluation to be made. Such a model therefore has a statistical nature and can be static. This is desirable due to the relative simplicity of statistical models over dynamical models.

For the sources of variations, we will mainly focus on the variations in inputs and that are imposed by processes. The choice for this is currently taken because the effect of external disturbances is not straightforward to characterize. We will however leave an option for the user to investigate the effect of these disturbances through our model / system analogy that will be presented later.

In summary; the central issues that we will pay attention to in this work is the question where measurements should be placed in a production line, so that a data-driven input-output model can be identified / verified. We will restrict ourselves to statistical and static model to describe the propagation of variations through the production line. We expect that such a model will be relatively straightforward to generalize. We do not expect to solve the measurement placement problem, since that requires highly accurate knowledge on all relations in a production line. Instead, we focus on what information is available when there are significant variations in a production line, and how this information can be used for decision-support on which measurements to necessary to be introduced.

4.6.3 Use Cases - Examples

The envisioned use cases concern production lines where sources of defects are difficult to be identified based solely on the operators' experience. The methodology that we propose can provide support in making smart decisions on where sensors can be placed to increase understanding on the output quality variations. The use case which we will consider explicitly during this development is that of the Philips Drachten QU4LITY pilot line.

4.6.4 Customization & Use in QU4LITY

The approach will be developed and validated in the QU4LITY project. No enhancements or customization from prior versions are therefore relevant. The contribution will be related to the Philips Drachten pilot line.

4.7 Analytics for In-Line Sensor Data Visualization

4.7.1 Overview

In the context of Additive Manufacturing (AM), the CAD technology 'Design4AM' provides new tools for intuitive volumetric modelling, mechanical simulation and interactive visualization of organically shaped parts. Thanks to an innovative data representation format and efficient data structures based on trivariate splines and volumetric subdivision techniques, we can handle even parts with properties varying in space like smooth material transitions inside the object [Altenhofen18]. The compact storage of information will then be exploited for visualization purposes enabling user interaction enriched with semantic computing techniques. Finally, our fast GPU-based simulation for structural analysis allows to check stability of such objects prior to production in few seconds – i.e. several factors faster than commercial software.

The whole concept from capturing spatially varying properties with our data structures until volumetric visualization via GPU-accelerated algorithms is not only restricted to smooth interior material transitions but rather any volumetric data sets that change with certain behavior inside the object. The application and customization of this workflow to in-line sensor data will be developed within the QU4LITY project.

4.7.2 Use Cases - Examples

Since organically shaped parts are commonly produced by AM (Additive Manufacturing) techniques and our underlying volumetric representation format is especially suited for this kind of geometry, the main field of application of our technology is the modelling and visualization of additively manufactured parts. Together with the interactive simulation aspect, this approach fits well into a workflow where topology-optimized objects need further adaptations by a designer interested in immediately experiencing the effects of the changes based on integrated simulation in the design & modelling front end.

Moreover, ongoing advancements in production techniques do and will enable the manufacturing of objects with smooth transitions between materials like metal or ceramics. Such graded objects promise to have better functional properties when compared to parts with abrupt material changes as increased stability due to elimination of potential breaking points or superior thermal conductivity and isolation [Gupta15]. For the efficient production of these functionally graded materials, the inner material distribution must be described in a fully volumetric and smooth way. Our technology 'Design4AM' perfectly addresses this issue providing a continuous description of material distribution.

From the ZDM point of view, our data structure and visualization strategy will be applicable to sensor data gathered during the printing process. Due to the existing use case in the PRIMA pilot, we will focus on an optical sensor system measuring the thickness or roughness of powder beds in a selective laser melting (SLM) process. The interactive visualization of massive sensor data is key to a human-interaction with in-line sensor data and facilitates the understanding and recognition of events occurring during manufacturing that potentially cause defects to the currently manufactured part. Additionally, the ability to visualize sensor data also in a post-processing setting enables learning effects of how to adjust process parameters in general and might even give feedback to a better object modelling step.

4.7.3 Customization for QU4LITY

Our technology is already able to represent material gradation (material particle distribution) inside the object. In QU4LITY, we will enhance this capability to also describe and approximate in-line sensor data gathered during the production of AM parts. Due to the layerwise production paradigm of AM, the volumetrically spread in-line sensor data, e.g. temperature measurements or optical sensor data measuring the thickness of the powder bed layer in a PBF machine, comes in slices arranged along the height of the object. Therefore, a customization to this special kind of data distribution is needed when we extend our technology to the approximation of in-line sensor data.

Since in-line sensor systems are expected to produce large data sets, the approximation approach with our innovative representation scheme will be key to achieve compression of these big sensor data. To maintain the ability to access the original sensor data exactly and error-free, the approximation approach will be complemented with data analysis methods like segmentation and selection. This combined approach will facilitate fast visualization with intuitive user interaction: with a visualization based on the compact representation of approximated data, the user may want to see the raw sensor data in a locally selected area. Thus, the original sensor data is only accessed partially and will be loaded on demand.

A possible workflow for the complete approximation and visualization pipeline for optical sensor data is depicted in Figure 14. Based on the sensor data given as PNG pictures of each slice and a STL file of the object model, the data compression and visualization algorithms operate on three different resolution levels: Fitting the whole sensor data set with a coarsely resolved, volumetric data structure gives a high compression and an overview of the whole data set. Selecting areas of interest will then give a zoom-in view of corresponding data layers that are again compressed by bivariate spline approximations with adaptive resolution. If the operator is still interested in the exact sensor data of a localized region, we may employ on demand loading of the requested segment from the raw data sets.

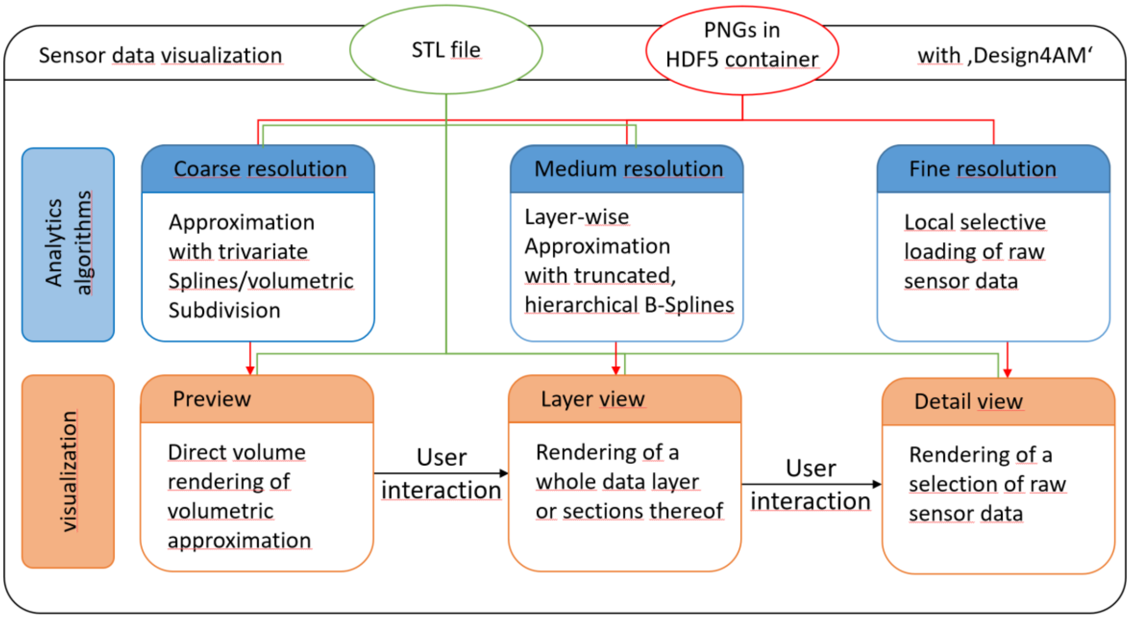


Figure 14: Planned workflow for visualization of in-line sensor data

4.7.4 Use in QU4LITY

The extension of 'Design4AM' to approximate in-line sensor data and the data segmentation analysis algorithm will be developed within the PRIMA pilot case where a powder bed fusion (PBF) machine is considered. In this production process, the correct powder bed thickness is essential for a flawless part since places with deviating thickness might result in unintended melting behavior and thus potentially causing defects in the object. The detection of such events is crucial in view of the ZDM context.

Therefore, another pilot partner, Fraunhofer ILT, will construct and integrate an optical sensor system into the PBF machine measuring the thickness of each powder bed layer. Based on these data, our visualization solution utilizing the approximation and segmentation approach will contribute to the detection of anomalies in the big sensor data. For a human-friendly and intuitive interaction, the sensor data will be visualized together with a CAD (Computer Aided Design) model of the to be printed object.

Since the development of our approximation and visualization workflow needs to be customized to the specifics of the PRIMA use case, the sensor system must be deployed beforehand such that the approximation workflow can be tested on the data sets. Therefore, the roadmap for the development is planned as follows:

- **M14:** First volumetric visualization of raw and uncompressed sensor data.

- **M17:** Approximation algorithm and an interactive graphical user interface for a coarse volumetric preview.
- **M23:** Approximation algorithm and interactive graphical user interface for a layer-wise and adaptively refined approximation.
- **M26:** Selective loading and rendering of raw sensor data segments.

5. Conclusions

Zero Defect Manufacturing processes in the scope of a digital shopfloor are always data intensive. Therefore, their deployment hinges on Big Data infrastructures and Big Data analytics algorithms. The use of certain classes of algorithms (notably predictive algorithms) is a key prerequisite for the implementation of the project's autonomous quality paradigm, which is based on the cognitive and proactive identification of quality issues as a means to remedying them in a timely fashion.

This deliverable has reported on the list of Big Data platforms and data analytics algorithms that are provided and used by QU4LITY partners. In particular, the deliverable has emphasized on presenting knowledge and IPR of the partners in terms of Big Data platforms and Big Data analytics algorithms, which will be bundled in the project's library of Big Data assets. This library will support multiple objectives :

- It comprises some of the main platforms and algorithms used in the project's pilots in WP7 of the project.
- It enable the validation of the Reference Architecture (RA) of the project (developed in WP2) based on practical platforms that will be used in the project's pilots.
- It provides a set of assets that will be included and made available through the market platform of the project in WP8.

In terms of platforms, the deliverable illustrates three Big Data platforms (i.e. FAR-EDGE Distributed Data Analytics Platform by AIT, Open VA platform by VTT, Data Fabric platform by LKS) that provide added-value features and are expected to accelerate the deployment of data-intensive systems in the ZDM pilots of the project. These platforms are currently under customization and enhancement in-line with the needs of the project pilots.

Likewise, this deliverable has presented a range of data analytics and machine learning algorithms (including deep learning) with clear relevance for ZDM and quality management problems. These algorithms span RUL calculation for the prediction of issues with the machineary, custom explainable rule-bases algorithms for the proactive identification of defects, as well as specialized algorithms for the optimal deployment of quality-related sensors and for the visualization of in-line sensor data.

Overall, the deliverable has reported on the Big Data infrastructure and analytics algorithms that will form the project's Big Data assets algorithms. The implementation of these infrastructures and their customization for the needs of the pilots is in progress. Their final implementation, including results on their performance on quality data will be presented in the second and final version of the present deliverable.

References

[Altenhofen18] Altenhofen C., Luu T.H., Grasser T., Dennstädt M., Mueller-Roemer J.S., Weber D. and Stork A.: Continuous property gradation for multi-material 3D-printed objects, Proceedings of the 29th Annual International Solid Freeform Fabrication Symposium – An Additive Manufacturing Conference. Austin, Tex.: University of Texas 2018, pp. 1675-1685

[Barrett74] Barrett, James P. "The coefficient of determination—some limitations." The American Statistician 28, no. 1 (1974): 19-20.

[Chakravarti67] Chakravarti, Laha, and Roy, (1967). Handbook of Methods of Applied Statistics, Volume I, John Wiley and Sons, pp. 392-394.

[Christou18] I.T. Christou et al. "A Parallel/Distributed Algorithmic Framework for Mining All Quantitative Association Rules", arXiv preprint arXiv:1804.06764, 2018

[Christou19] I.T. Christou, "Avoiding the Hay for the Needle in the Stack: Online Rule Pruning in Rare Events Detection", Proc. IEEE ISWCS 2019, Oulu, Finland, Aug. 27-29, 2019.

[D'Agostino86] D'Agostino, Ralph B. Goodness-of-fit-techniques. Vol. 68. CRC press, 1986.

[Gelman03] Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. "Bayesian Data Analysis. Chapman & Hall/CRC." In :. 2003.

[Lu97] J. C. Lu, J. Park, Q. Yang, "Statistical inference of a time-to-failure distribution derived from linear degradation data", Technometrics, vol. 39, pp. 391-400, 1997.

[Luong17] M.-T. Luong et al. "Effective Approaches to Attention-based Neural Machine Translation", https://nlp.stanford.edu › pubs › emnlp15_attn, 2017.

[Motulsky87] Motulsky, Harvey J., and Lennart A. Ransnas. "Fitting curves to data using nonlinear regression: a practical and nonmathematical review." The FASEB journal 1, no. 5 (1987): 365-374.

[Myung03] Myung, In Jae. "Tutorial on maximum likelihood estimation." Journal of mathematical Psychology 47, no. 1 (2003): 90-100.

[Peng09] Peng, Chien-Yu, and Sheng-Tsaing Tseng. "Mis-specification analysis of linear degradation models." IEEE Transactions on Reliability 58, no. 3 (2009): 444-455.

[Sakamoto86] Sakamoto, Yosiyuki, Makio Ishiguro, and Genshiro Kitagawa. "Akaike information criterion statistics." Dordrecht, The Netherlands: D. Reidel 81 (1986).

[Sherstinksi18] A. Sherstinksi, "Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) Network", arXiv:1808.03314 [cs.LG], 2018.

[Si11] Si, Xiao-Sheng, Wenbin Wang, Chang-Hua Hu, and Dong-Hua Zhou. "Remaining useful life estimation–a review on the statistical data driven approaches." European journal of operational research 213, no. 1, 1-14, 2011.

[Snedecor89] Snedecor, George W. and Cochran, William G. (1989), Statistical Methods, Eighth Edition, Iowa State University Press.

[Stephens74] Stephens, M. A. (1974). EDF Statistics for Goodness of Fit and Some Comparisons, Journal of the American Statistical Association, 69, pp. 730-737.

List of figures

List of tables

List of Abbreviations

Acronym	Abbreviation
AI	Artificial Intelligence
AM	Additive Manufacturing
CNN	Convolutional Neural Network
CAD	Computer Aided Design
CSC	Cyber System Connector
CTQ	Critical To Quality
DDA	Distributed Data Analytics
DL	Deep Learning
DSS	Decision Support System
ETL	Extract Transform Load
GUI	Graphical User Interface
HMMs	Hidden Markov Models
IoT	Internet of Things
LSTM	Long Short-Term Memory
ML	Machine Learning
PBF	Powder Bed Fusion
QARMA	Quantitative Association Rule Mining
RA	Reference Architecture
RNN	Recurrent Neural Networks
RUL	Remaining Useful Life
ZDM	Zero Defect Manufacturing

Partners:

INTELLIMECH CONSORZIO PER LA MECCATRONICA	SIEMENS	AtoS
AIRBUS	MONDRAGON	NXT CONTROL
ATB Institut für angewandte Systemtechnik Bremen GmbH	Continental	RIA STONE
UNIMETRIK METROLOGY AND CALIBRATION	Technology Transfer System	EPFL
SQS	FAGOR FAGOR ARRASATE	TTTech Ensuring Reliable Networks
tu technische universität dortmund	thyssenkrupp	Jožef Stefan Institute Ljubljana, Slovenia
PRIMA INDUSTRIE	eit Digital	ENGINEERING
PACE a TXT company	innovalia ASSOCIATION	Telefónica
Synesis sustainable automation	ghi	UNPARALLEL
DANOBAT	POLITECNICO MILANO 1863	Whirlpool CORPORATION
VTT	VISUAL COMPONENTS	ASTI
Fraunhofer	cea	INDUSTRIAL DATA SPACE ASSOCIATION
PHILIPS	ATLANTIS	aic AUTOMOTIVE INTELLIGENCE CENTER
KOLEKTOR	AIT	SINTEF
+GF+	Technische Universität Braunschweig	TNO innovation for life