D4.9 – i4Q Data Integration and Transformation Services v2

WP4 – BUILD: Manufacturing Data Analytics for Manufacturing Quality Assurance

i4Q has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 958205.

Document Information

GRANT AGREEMENT NUMBER	958205	ACRONYM	i4Q
FULL TITLE	Industrial Data Services for Quality Control in Smart Manufacturing		
START DATE	01-01-2021	DURATION	36 months
PROJECT URL	https://www.i4q-project.eu/		
DELIVERABLE	D4.9 – i4Q Data Integration and Transformation Services v2		
WORK PACKAGE	WP4 – BUILD: Manufacturing Data Analytics for Manufacturing Quality Assurance		
DATE OF DELIVERY	CONTRACTUAL 31-Dec-2022	ACTUAL	30-Dec-2022
NATURE	Report	DISSEMINATION LEVEL	Public
LEAD BENEFICIARY	CERTH		
RESPONSIBLE AUTHOR	Athina Tsanousa (CERTH), Ilias Gialampoukidis (CERTH)		
CONTRIBUTIONS FROM	-		
TARGET AUDIENCE	1) i4Q Project partners; 2) industrial community; 3) other H2020 funded projects; 4) scientific community		
DELIVERABLE CONTEXT/ DEPENDENCIES	This document presents the data integration and transformation services (version 2) within i4Q. This document has a preceding version, namely D4.1.		
EXTERNAL ANNEXES/ SUPPORTING DOCUMENTS	None		
READING NOTES	None		
ABSTRACT	Data integration and transformation services is a server-based solution responsible for delivering datasets ready for further analysis. The current deliverable summarizes the progress done for the second version of the solution.		

Document History

VERSION	ISSUE DATE	STAGE	DESCRIPTION	CONTRIBUTOR
0.1	07-Nov-2022	ToC	Table of contents creation	CERTH
0.2	25-Nov-2022	Draft	1st draft available for internal review	CERTH
0.3	02-Dec-2022	Internal Review	Internal review	UNINOVA, IKERLAN
0.4	09-Dec-2022	Draft	2nd Draft addressing comments from internal reviewers	CERTH
1.0	30-Dec-2022	Final Document	Final quality check and issue of final document	CERTH

Disclaimer

Copyright message

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS/ACRONYMS

DIT	Data Integration and Transformation
DR	Data Repository
DSS	Decision Support System
EOL	End Of Line
FFT	Fast Fourier Transformation
LGBM	Light Gradient-Boosting Machine
SMOTE	Synthetic Minority Oversampling Technique
UI	User Interface

Executive summary

This document presents an executive explanation of the **i4Q Data Integration and Transformation Services** (i4Q^{DIT}) Solution, version 2, providing the general description, the technical specifications and the implementation status. This document i4Q D4.9 v2 is an update of v1 of D4.1., for this reason it contains information of the 1st version together with the updates developed in this 2nd version.

The documentation associated to the i4Q^{DIT} Solution is deployed on the website http://i4q.upv.es. This website contains the information of all the i4Q Solutions developed in the project "Industrial Data Services for Quality Control in Smart Manufacturing" (i4Q). The direct link to the i4Q^{DIT} Solution documentation is https://i4q.upv.es/9_i4Q_DiT/index.html

Such documentation is structured according to:

- General description
- Features
- Images
- Authors
- Licensing
- Pricing
- Installation requirements
- Installation Instructions
- Technical specifications of the solution
- User manual

Document structure

Section 1: Contains a general description of the **i4Q**DIT, providing an overview and the list of features. It is addressed to final users of the i4Q Solution.

Section 2: Contains the technical specifications of the **i4Q**DIT, providing an overview and its architecture diagram. It is addressed to software developers.

Section 3: Details the implementation status of the **i4Q**DIT, explaining the current status, next steps and summarizing the implementation history.

Section 4: Provides the conclusions.

APPENDIX I: Provides the PDF version of the **i4Q**DIT web documentation, which can be accessed online at: **https://i4q.upv.es/9_i4Q_DiT/index.html**

1. General Description

1.1 Overview

i4Q^{DIT} is a distributed server-based solution able to prepare manufacturing data for being efficiently processed by other analytical solutions. The functions that are included in i4Q^{DIT} and are required for manufacturing data stream management are: reading, cleaning, storing, indexing, enriching, searching & retrieving, maintaining, and correspondence of open APIs. The solution offers a variety of pre-processing steps, such as filtering, decomposition and feature extraction, that transform the complex raw data received from manufacturing processes, into suitable formats for further exploitation.

This version of the deliverable is an updated one of the D4.1. This version includes the functions developed for the rest of the pilots, as well as some information about the user interface developed for three solutions: i4Q^{DIT}, i4Q^{IM}, and i4Q^{QD}.

1.2 Features

i4Q^{DIT} solution consists of the following features some which have already been mentioned in D4.1:

1. Reading of different types of data produced by the manufacturing workflows.

3. Connect with Data Repository (DR) in order to receive raw data and then send the pre-processed data, ready for the analysis.

4. Data preparation functions: cleaning, pre-processing, feature extraction, data filtering, data harmonization.

5. Provide integrated datasets with a homogenous structure of different recordings from multiple sources, so as to prepare the manufacturing data for further analysis.

In this update version of the solution, more pre-processing functions have been developed, such as the SMOTE (Synthetic Minority Oversampling Technique) method for creation of synthetic data in unbalanced samples.

2. Technical Specifications

2.1 Overview

i4Q^{DIT}, or Data Integration and Transformation Services, is mapped to the Platform Tier but mostly operates on the Edge Tier. This solution focuses mainly on data transformation at the Platform Tier, where data are prepared for processing by microservice applications. Edge Tier services focus primarily on data collection, management, and analysis. This solution's inputs consist of sensors and other sorts of manufacturing data that require integration and pre-processing. The output consists of databases, both online and offline, containing clean, integrated data that will be given to other components for additional processing and analysis.

2.2 Architecture Diagram.

The processes and services that are being included in the i4Q^{DIT} software tool are mapped to two tiers in the i4Q Reference Architecture, Platform and Edge Tier (**Figure 1**). i4Q^{DIT} is one of the most commonly used solutions in the i4Q and for that reason, some strengths and weaknesses arise.

- **Platform Tier:** "Data Transformation" is one of the most important service of this solution, as it is responsible for transforming the data (usually post-processing) for each individual need for the microservices that require them. The main benefit of this service is that it can obtain and use data from repositories, already stored and ready to be transformed to a usable form. This gives the ability to use this solution without the need for newly acquired data.
- **Edge Tier:** i4Q^{DIT} is the main solution responsible for the transformation and integration of data. Such operation requires services such as "Data Collecting", "Data Management", "Data Services" and "Distributed Computing". "Distributed Computing" provides a model in which components of the software tool are shared among multiple computers (nodes) that allow deploying and running AI workloads on the edge. The benefit of this service is that the solution can be used on the manufacturing floor. "Data Collecting", "Data Management" and "Data Services" can all be considered part of a pipeline, from data ingestion (collecting raw data from the facilities and storing them to make them available for pre-processing), to data management and transformation (pre-processing the data so that they can be fit to be used by other solutions). The use of these services is crucial for the proper development and operation of this solution.

  i4Q^{DIT} is mapped to "Data transformation" sub-component of the Platform Tier and will be using:

  - Data analytics and services: cooperate with this solution to provide integrated datasets for further analysis.
  - Data management: useful for all data driven solutions.
  - Data services: to send the integrated and fused datasets.
  - Data in motion: to draw real time data for further pre-processing and integration.
  - Data at rest: to extract data for further pre-processing.

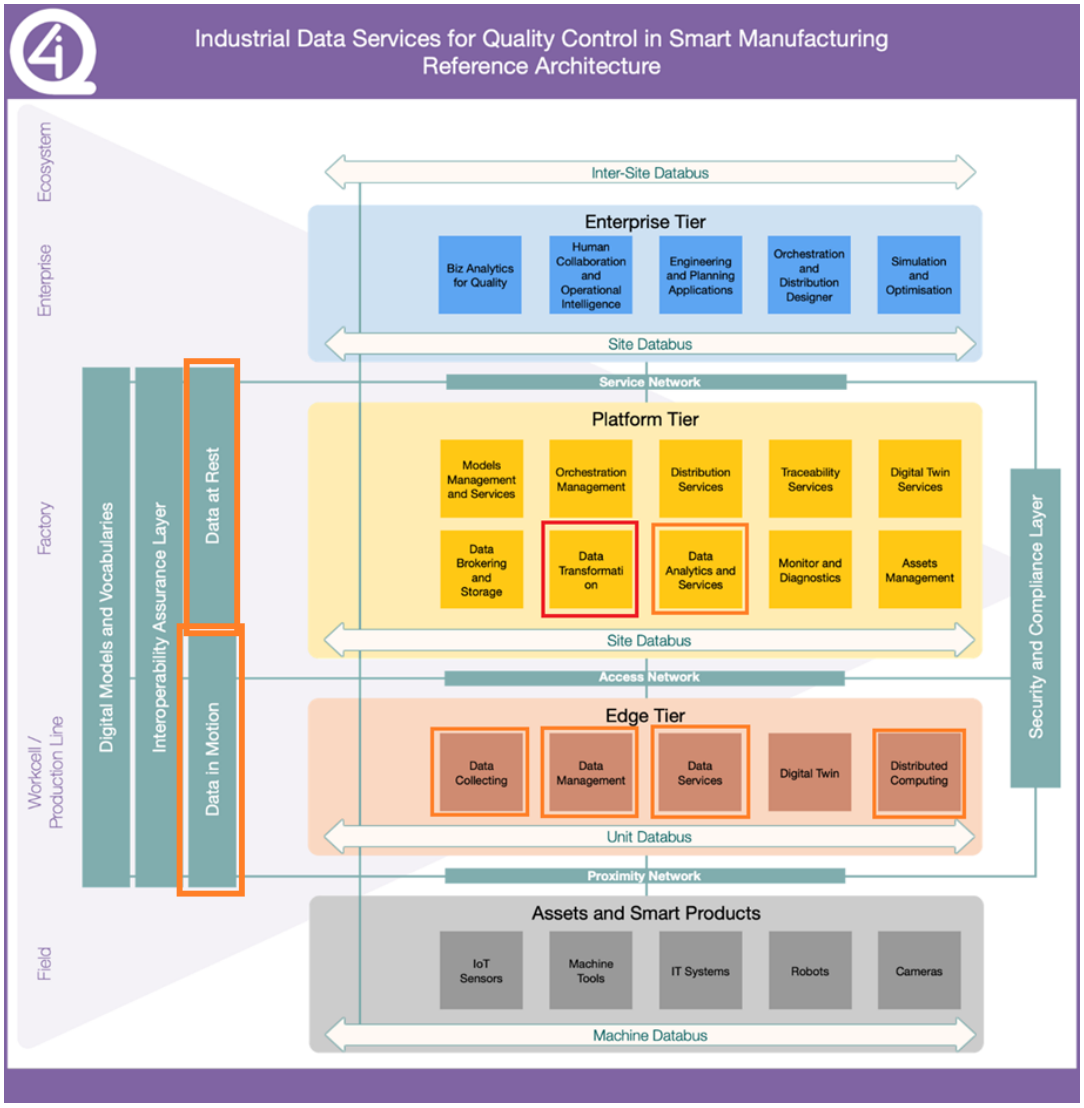

Figure 1. Reference Architecture

3. Implementation Status

3.1 Current implementation

3.1.1 Solution features analysed and mapping with user requirements

The current developments of the solution do not map to any of the requirements described in WP1, but they cover the requests of the pilots. Among the already existing requirements, the ones that have not yet been addressed by i4QDIT, refer to real time applications, which are not yet available. The state of the pilots determines the future development of real-time applications.

3.2 New developments included in v2

3.2.1 Imbalanced datasets

The problem of imbalanced classes is one that can frequently be encountered in datasets that are being used for classification, especially in real world applications from the industrial sector, where the amount of faulty products is usually quite small. Data imbalance normally occurs as an uneven distribution of classes throughout a dataset (**Figure 2**). If a binary classification model is trained without correcting this issue, the model will be entirely biased after training is complete. It also has an effect on the correlations that exist between the features.

The following is an overview of some of the solutions that can be used for the class imbalance problem:

- Undersampling is the process of arbitrarily removing some of the observations from the class that constitutes the majority in order to bring those numbers closer with those of the class that constitutes the minority.
- Oversampling is the way of referring to the second type of resampling method. Undersampling isn't as complicated as this technique, yet it's still rather complex. The technique of producing synthetic data involves making an attempt to generate at random a sample of the features based on the observations of the minority class. When dealing with a standard classification issue, one can oversample a dataset using any one of a number of different approaches. The most typical approach is referred to as SMOTE (Synthetic Minority Over-sampling Technique). To put it in simple terms, it examines the feature space that contains the minority class data points and takes into account the k data points that are located closest to it.

The **SMOTE** method has been developed and implemented on a few pilot cases (Factor and Farplas). Specifically, in Farplas's pilot case the initial data were comprised of multiple classes and the dataset had to be transformed into a binary dataset. The transformation from multiclass to binary problem was important due to the major imbalance in the dataset. The majority class has around 100k samples while the largest of the minority has around 800 samples. By converting the problem to binary and fusing all the faulty classes as one the dataset remains imbalanced but the ratio (**Class=1, n=108067 samples (98.203%) Class=0, n=1978 samples (1.797%)**) has now changed and resampling techniques can be applied.

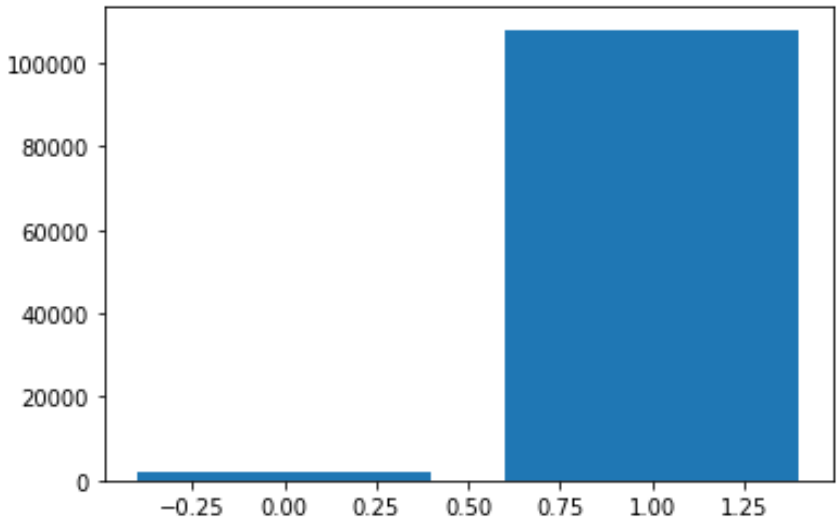

Figure 2. Imbalanced dataset

After applying SMOTE (**Figure 3**) and using a ratio of 30% the dataset transforms as follows (**Class=1, n=108068 samples (76.923%) Class=0, n=32420 samples (23.077%)**):

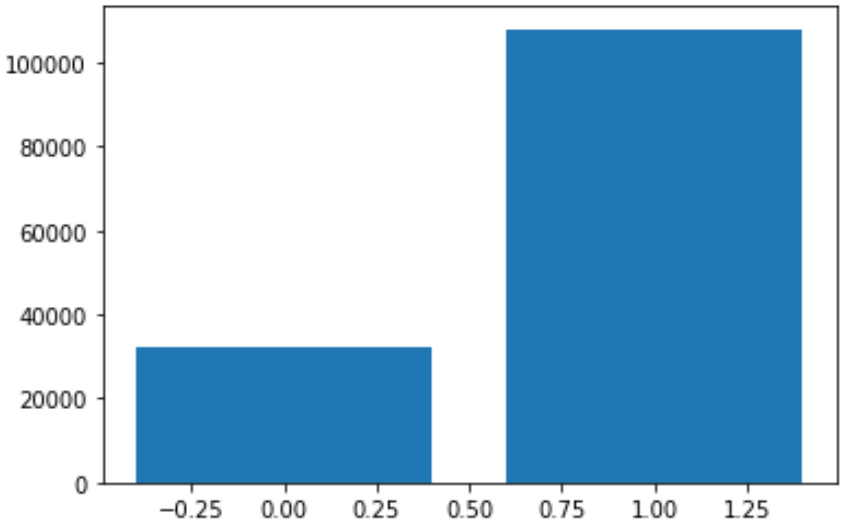

Figure 3. Dataset after applying SMOTE

After transforming the dataset i4Q[QD] utilizes it to classify the quality of the parts.

Functions of the solution were also developed for Whirlpool's pilot. A collection of data was provided by Whirlpool, with each one requiring different type of analysis. The description of data and the work done can be found in the following sub-sections.

3.2.2 End Of Line (EOL) data:

The EOL dataset (WHIRLPOOL) is built by testing several components of a product and evaluating the quality of those components. The dataset can be broken down into four distinct categories: Header, Values, Variables, Production. The information pertaining to the testing procedure can be found in each of the individual files. The information on the time and date of the test is included in the Header file, together with product data (the serial number), production data (the line and place id), and the test results. The timestamps and measurements that were taken in the course

of the testing are included in the Values file (voltage, power, current). Once more, the timestamps, measurement identifiers, and results can all be found in the Variables file. Last but not least, the Production file was the final one that was received. This file includes data that relates to the Serial Number of the product as well as the materials that were utilized along with other unique ids.

The size of the data is around 6 gigabytes, and the essential procedures were carried out to guarantee that i4Q^{DIT} and the scripts that were written could deal with files of this size without resorting to the use of big data approaches and technologies.

Due to their reasonable size, the Headers and Variables were merged using their respective unique ids. The Values file is where all of the sensor readings from the various testing methods are stored. Because the sensors collected data at such a high frequency, many of the measurements that were taken were redundant. As a result, the scripts that were processing the file estimated the mean values and changed the dataset into a much more manageable and compact form. Following the completion of that process, the datasets were combined while retaining their own unique identifiers.

After merging the first three files, some scripts were utilized in order to conduct an analysis of the dataset. In addition to correlations between the various data fields, missing data analysis and descriptive analytics were also carried out. Although the discovered correlations were not very strong, the dataset produced some very impressive findings when combined with an LGBM (light gradient-boosting machine) model obtained from the i4Q^{QD} solution.

In the end, the Production dataset was developed in order to differentiate the products based on the serial number, the material that was being used, and a variety of other data fields. Work is still being done to complete the analytics utilizing the new dataset.

Python, scikit learn, dtale, missingno, pandas, numpy, and *matplotlib* are some libraries that were utilized during the process of developing those scripts.

3.2.3 SPC data:

SPC (Statistical Process Control) dataset is a collection of critical measurements regarding parameters of an appliance's performance. The requirement of the pilot (Whirlpool) was to explore different groups of data and extract useful knowledge. For this reason, pre-processing steps and data analytics were developed. functions of the solution, developed for this dataset, were suitable for removal of wrong entries, merging of data according to a key variable, creation of plots such as boxplots and functions for checking if the values of the variable of interest are inside desired limits.

3.2.4 User interface

The UI (user interface) that has been developed consists of three components: the header, the side menu and the main part which contains the chart (**Figure 4**). The header includes the i4q project's logo and title ("Industrial Data Services for Quality Control in Smart Manufacturing"). The side menu on the left of the UI allows a user to select one of the existing use cases: FFT (Fast Fourier Transformation), Outliers, Chatter or Degradation. The first two cases (FFT and Outliers) correspond to existing functions of the v1 of the i4Q^{DIT}, while the other two cases correspond to solutions provided by CERTH. When a user selects an option, the corresponding chart is displayed

in the main part of the UI. There is also the ability to zoom in and out the chart in order to focus on a specific part of it. The data values for each of the four use cases are fixed and stored in a CSV file respectively. When a use case is selected from the menu, the corresponding CSV file is read on the server, the data are sent to the front-end where they are formatted accordingly and passed as input to the chart that displays them.

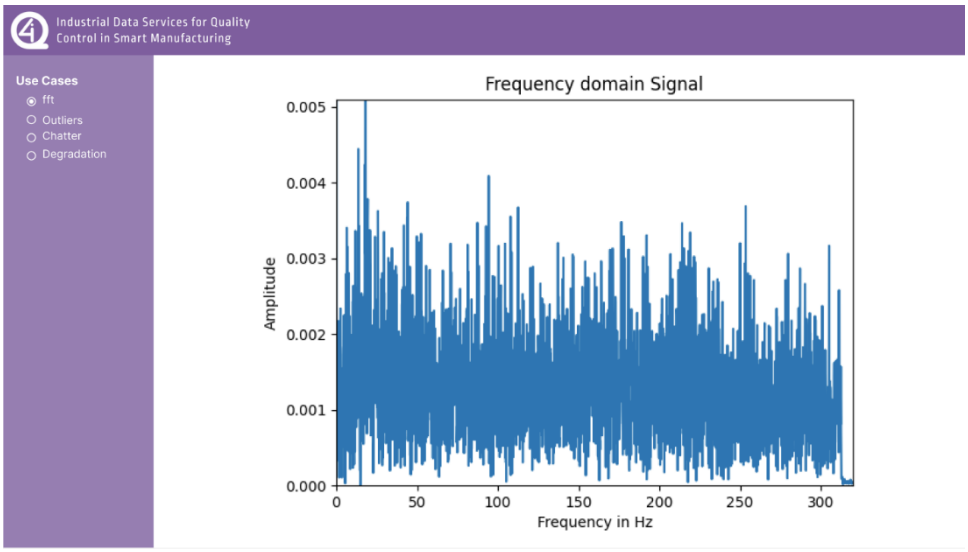

Figure 4. User Interface example

The technologies that have been used for the development of the UI are HTML5, CSS3 and JavaScript. A JavaScript library named Chart.js (https://www.chartjs.org/) is used for creating the charts. The zoom feature is implemented by the Chart.js plugin, chartjs-plugin-zoom (https://www.chartjs.org/chartjs-plugin-zoom/latest/).

3.3 History

Version	Release date	New features
V0.1.0	7/11/2022	TOC
v1	15/11/2022	Adding input
v2	25/11/2022	Version for internal review

Table 1. i4QDIT History

4. Conclusions

This deliverable summarizes the newest developments in the second version of the solution "Data integration and transformation services". The solution has already covered requirements from the following pilots: FIDIA, WHIRLPOOL, FARPLAS, FACTOR, BIESSE. In the upcoming months, i4Q^{DIT} will be adapted in the generic pilot and will develop the real time features if requested from the pilots. Finally, the user interface for i4Q^{DIT} and the rest of solutions provided by CERTH, will be delivered.

Appendix I

i4Q < Data Integration and Transformation Services> web documentation can be accessed online at: **http://i4q.upv.es/9_i4Q_DiT/index.html .**