Ref. Ares(2020)5091765 - 29/09/2020

FACTLOG
www.factlog.eu

ENERGY-AWARE FACTORY ANALYTICS FOR PROCESS INDUSTRIES

Deliverable D2.1

Analytics System Requirements and Design Specification

Version Version 1.1	**Lead Partner** JSI
Date 28/09/2020	**Project Name** FACTLOG – Energy-aware Factory Analytics for Process Industries

FACTLOG has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement Nº 869951

Call Identifier H2020-NMBP-SPIRE-2019	**Topic** DT-SPIRE-06-2019 - Digital technologies for improved performance in cognitive production plants
Project Reference 869951	**Start date** November 1st, 2019
Type of Action IA – Innovation Action	**Duration** 42 Months

Dissemination Level

X	**PU**	Public
	CO	Confidential, restricted under conditions set out in the Grant Agreement
	CI	Classified, information as referred in the Commission Decision 2001/844/EC

Disclaimer

This document reflects the opinion of the authors only.

While the information contained herein is believed to be accurate, neither the FACTLOG consortium as a whole, nor any of its members, their officers, employees or agents make no warranty that this material is capable of use, or that use of the information is free from risk and accept no liability for loss or damage suffered by any person in respect of any inaccuracy or omission.

This document contains information, which is the copyright of FACTLOG consortium, and may not be copied, reproduced, stored in a retrieval system or transmitted, in any form or by any means, in whole or in part, without written permission. The commercial use of any information contained in this document may require a license from the proprietor of that information. The document must be referenced if used in a publication.

Executive Summary

This document describes in detail all the requirements for the analytics system to be set-up and used in the FACTLOG project. The requirements are collected and presented on a per-pilot basis, considering the types and volumes of data available for each pilot as well as the target problems. Finally, the design specifications for the analytics system are described, both from the conceptual and the technical standpoint.

Each pilot use-case has a set of target problem scenarios they want to address with the technology from the FACTLOG platform. This deliverable presents all these scenarios for each pilot and identifies the role of analytics in them. Once each scenario is formulated as an analytics problem the methodology for addressing it is identified. The data sources and types are also inspected, and it is assessed if they are appropriate for the planned approach. An aggregated overview of the requirements is given for clarity.

Based on the requirements a design specification for the analytics system is drafted. First the analytics system is placed in relation to other components. It's role as a building block of the cognitive factory framework is explained as well as the interactions it has with the optimisation system and the knowledge graph and process models. Then, a set of tools (i.e. analytics libraries and platforms) is identified along with the methods and approaches that address the requirements collected in the preceding sections of the document.

The deliverable is a comprehensive collection of requirements for the analytics system and the specification of its conceptual and technical design. It is built on the information and insights collected regarding the pilots and the project challenges up to the time of its preparation. The requirements and specifications may evolve as the project progresses and any adaptations will be reported in future deliverables.

Revision History

Revision	Date	Description	Organisation
0.1	17/06/2020	Table of contents	JSI
0.2	17/08/2020	Initial draft of Section 3 - Analytics System Design	JSI
0.3	21/08/2020	Description of JEMS requirements and data in Section 2.1	QLECTOR
0.4	22/08/2020	Description of the TUPRAS requirements and data in Section 2.1	JSI, TUPRAS, MAG
0.5	23/08/2020	Description of the BRC requirements and data in Section 2.1	JSI, C2K, BRC, AUEB, UNIPI
0.6	15/09/2020	Description of the Continental requirements and data in Section 2.1	JSI, SIMAVI, CONT
0.7	18/09/2020	Description of the Piacenza requirements and data in Section 2.1	JSI, PIAC, SIVECO
1.0	21/09/2020	Final sections	JSI
1.1	28/09/2020	Incorporated comments from the internal review.	JSI, NISSA, TUC

Contributors

Organisation	Author	E-Mail
JSI	Aljaž Košmerlj	aljaz.kosmerlj@ijs.si
QLEC	Klemen Kenda	klemen.kenda@qlector.com
QLEC	Jože Rožanec	joze.rozanec@qlector.com
TUPRAS	Melike Kamuran Onat	melike.onat@tupras.com.tr
AUEB	Stavros Lounis	slounis@aueb.gr
AUEB	Georgios Zois	georzois@aueb.gr
AUEB	Yiannis Mourtos	mourtos@aueb.gr
UNIPI	Pavlos Eirinakis	pavlose@unipi.gr
UNIPI	Konstantinos Kaparis	k.kaparis@uom.edu.gr
DOM	Caterina Calefato	Caterina.calefato@domina-biella.it
DOM	Marco Vallini	Marco.vallini@domina-biella.it
PIA	Eugenio Alessandro Canepa	alessandro.canepa@piacenza1733.it
C2K	Kevin Greening	kgreening@control2k.co.uk
BRC	Alexander Adams	alexander.adams@brc.ltd.uk
SIMAVI	Andrea Paunescu	andreea.paunescu@siveco.ro
CONT	Alin Popa	alin.3.popa@continental-corporation.com

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Purpose and Scope

This document collects the requirements for the analytics system from the pilot cases and aggregates them into a comprehensive list that will guide the development from both the methodological and technical perspective. Requirements collection took into account both the operational needs of the pilots as well as the data availability. The operational needs determine what kind of results are required by each pilot to improve their production flow. This dictates the capabilities the analytics services need to offer. The data further informs the technical aspects of the services: what type and format of data they need to be able to consume; what volume of data they need to be able to process; and if there are any pre-processing steps needed before inputting the data into the analytics algorithms.

The analytics services are typically the first step in the data processing path. They clean the data and distil higher-level signals from it. For example, the analytics system may detect an upcoming machine malfunction in a predictive maintenance scenario. This prediction is a signal that is used, along with any relevant context data, to activate other systems such as optimisation. These subsequent systems then take in the outputs of analytics as inputs to perform their functions – e.g. in this case schedule maintenance as soon as possible in such a way as to minimally disrupt the production process and the delivery deadlines.

After the requirements are collected and aggregated, an overview of methods and tools that address the requirements is given. The overview presents a selection of approaches that can satisfy the pilots' needs as they are understood at the time of preparation of this document. The best options will be selected, and other methods may be added as the project progresses and more insights are gained into the pilots and their data.

It is important to note, that at the time of preparation of this deliverable, not all the datasets were available to the technical partners for detailed study. The legal details of sharing the data among the partners (including preparation and signature of the non-disclosure agreements) had to be resolved first. The technical aspects of sharing the datasets in an efficient manner also took time. These processes were foreseen, but their duration was significantly prolonged by the COVID-19 outbreak. These processes involved a lot of people from different departments (i.e. legal, management, archive…) and were significantly impacted by closures. Specifically, Piacenza and Continental factories closed a month and a half and many employees, including those directly involved in the FACTLOG processes, were put in layoffs in order to recover some costs. Consequently, this deliverable has been delayed and its contents are partially based on data specifications and samples, rather than full datasets.

Despite the drawbacks listed in the previous paragraph, the technical partners in charge of analytics are confident the requirements of the pilots had been addressed well. The consortium has a wealth of previous experience to lean on and is able to anticipate the specifics of individual cases. Furthermore, part of the work plan of some pilots (such as Piacenza and BRC) was the construction and expansion of their data infrastructure, so the project was prepared for adaptation to developing data needs from the start of the project. Any and all changes or deviations from the designs in this document will be reported in following deliverables.

1.2 Relation with other Deliverables

This deliverable builds on top of the descriptions of the pilot use-cases in D1.1 Reference Scenarios, KPIs and Datasets. Though short summaries of the pilots are given in sections 2.1 to 2.5, a reader should consult D1.1 for their full descriptions.

This document serves as a base for a lot of the future work in the project. It contains the requirements specification for all the next steps within WP2 and is as such the base for all its future deliverables: D2.2 Analytical Platform for Process Industry, D2.3 Holistic Model of Uncertainty and Causal Relations, D2.4 Anomaly Detection System, D2.5 Manufacturing Chain Recommender System.

The analytics system, the requirements of which are detailed in this document, is a building block of the cognitive factory framework together with other components. An overview of that role is explained in section 3.1.1 – the framework will be described in detail in deliverable D1.2 Cognitive Factory Framework. The technical specifications listed in the section 3.2 are going to be integrated into the specifications for the entire Factlog system in deliverable D1.3 System Architecture and Technical Specifications.

The services and tools built based on the requirements and specifications in this document will be described in deliverable D3.2 Data Analytics as a Cognitive Service. Those services and tools that focus on methods that make use of the structured knowledge in knowledge graphs prepared in WP4 will be described in D4.4 KG-based Analytics for Process Optimization.

Finally, since analytics are one of the main consumers of data in the Factlog system, their requirements and specifications are an important input for the design of the data acquisition and transfer systems detailed in D6.1 and D6.2 Data Collection Framework.

1.3 Structure of the Document

The introduction first outlines what is the purpose of this document and the areas if covers in section 1.1 and then lists all the deliverables related to this document, serving either as its input or representing future work based upon its content, in section 1.2.

The requirements section follows with individual pilots' requirements as well as their data types and sources explained in subsections – namely JEMS in 2.1, Tupras in 2.2, Piacenza in 2.3, Continental in 2.4 and BRC in 2.5. An aggregation of the requirements over all pilots is then given in section 2.6.

The next major section describes the design of the analytics system. First, the conceptual design is presented in 3.1 which also places the analytics system in the relation with the other components and work packages in 3.1.1. The technical design of the analytics system follows in section 3.2, with the list of tools planned for use in section 3.2.1 and the list of methods that address the pilots' requirements in section 3.2.2.

Finally, the appendices hold the detailed tables of meta-information about the datasets. There is one appendix per pilot, namely: Appendix I – JEMS Data, Appendix II – Tupras Data, Appendix III – Piacenza Data, Appendix IV – Continental Data and Appendix V – BRC Data.

2 Analytics requirements

This section presents the analytics requirements of all the FACTLOG pilots. The detailed descriptions of the pilots can be found in deliverable D1.1, here the focus is on identifying the roles of analytics in the pilot scenarios and formulating the methodology and the technology needed to fulfil those roles.

2.1 Waste to Fuel Transformer Plants: Pilot Case by JEMS

JEMS is developing waste-to-fuel transformer plants. The plants transform hydrocarbon-based waste into synthetic diesel fuel through a chemical de-polymerization process. An overview of the process flow is shown in Figure 1. The plants are designed for continuous operation and run the process in a multi-stage pipeline from input to output. The challenge in the JEMS FACTLOG pilot is to use cognitive digital twin technology to ensure the optimal operation of the plant and to avoid malfunctions. The plant is equipped with a large number of sensors monitoring its operation in real time and provide the data to facilitate the cognitive twin.

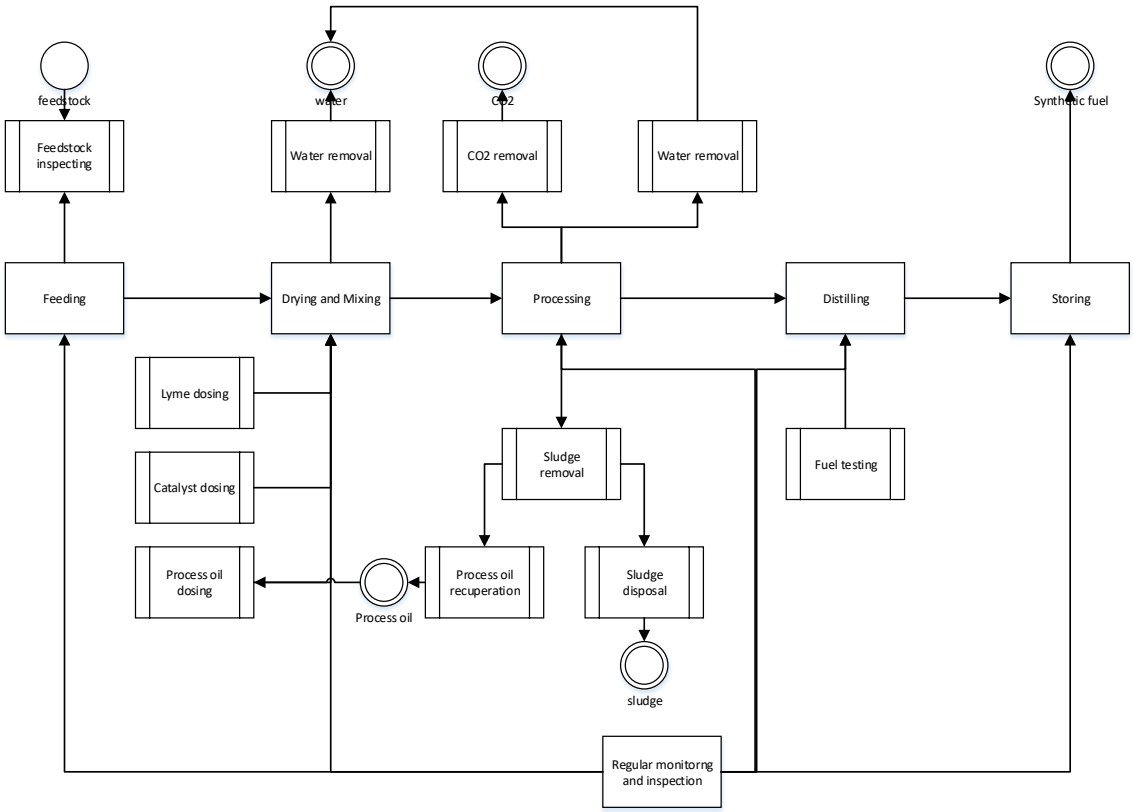

Figure 1: The JEMS plant operating process flow

2.1.1 Problem Scenarios and Requirements

There are two scenarios in the JEMS pilot.

Scenario #1: Clogging of pipes

Description: The material being processed in the plant passes through the plant pipes from the feeding stage at the start all the way to the output after the final distillation (the central

boxes in the diagram in Figure 1 with arrows from Feeding to Storing). At input the waste is ground down into small pieces and mixed with processing oil so it can pass through the pipes. Since the material is still quite dense and non-homogeneous, the pipes can get clogged. Cleaning the pipes can stop production for up to a day and is therefore costly. If clogging is detected beforehand, there are several actions available to prevent it (e.g. filtering or adding more oil) depending on the cause of the clogging.

Analytics approach: This scenario represents an anomaly detection problem. When a pipe is getting clogged the values reported by the plant sensors deviate from those recorded during normal operation. This can be detected by observing discrepancies between the sensor values and those generated by the simulation based on historic values. The root cause of the anomaly, which in this case means which pipe exactly is getting clogged and why, can be identified either by using process models built by experts or by using a classification model which finds malfunctions of the same type in historic data.

This is a general anomaly detection approach that can detect any kind of anomaly, not just clogging of the pipes. If its performance would prove to not be satisfactory, a targeted classification model could be built using stream learning methodology by labelling the target clogging situations in the data. The general and targeted approach can run side by side with the targeted model catching the known critical failures, while the general detector covers any other problems.

Scenario #2: New input materials

Description: The plant can process any type of hydrocarbon-based waste, which includes for example old wood, garden trimmings or even plastic garbage. These types of waste differ significantly in their properties such as calorific value, water content, chunk size etc. The plant operating parameters need to be set appropriately to ensure optimal processing of the input materials and determining the best parameter set can be a slow process that takes days. Since the plant is designed for continuous operation, it would ideally be able to automatically adapt to the new material or even variations in the same input material batch.

Analytics approach: To explore the space of possible parameters without actually running the plant we need to be able to simulate the plant operation. We can achieve this by modelling the plant stages taking the parameters and sensor values on the inbound pipes as input data and predicting the sensor values on the outbound pipes. A streaming regression model or an artificial neural network are models that can achieve this purpose. The exploration of the parameter space can be formulated as an optimisation problem for the optimisation component of the FACTLOG platform, however a reinforcement learning approach is also applicable as a solution from the field of machine learning.

2.1.2 Data Types and Sources

The chief source of data in the JEMS case are the sensors from the waste processing plant. The entire process is monitored from input to output and the data includes machine state values such as motor speeds or valve open/closed status as well as operations measurements such as temperature or pressure. Besides being used for monitoring the values are also stored in a historian database. The list of values along with their properties is given in Table 2 included in Appendix I – JEMS Data.

Note that the table contains more than 170 values, which is the number of sensors reported in deliverable D1.1. JEMS is actively working on extending the sensor array and the table contains some sensors which are new and whose values are not present in the current data samples. As the analytics will eventually need to handle the full set and the new sensors are similar in type than the old ones, we report the full current set in Appendix I – JEMS Data.

Roughly two years' worth of historical data is available from 22.6.2015 to 21.1.2017 amounting to about 10 Gb of data. All the values are either real-valued numbers or Boolean values (true/false). Both the data volume and the data types are well suited for the analytics algorithms and models referenced in section 2.1.1.

2.2 Oil Refineries: Pilot Case by TUPRAS

The Tupras oil refinery processes raw oil into several petroleum products such as Liquified Petroleum Gas (LPG), naphtha, gasoline, diesel and fuel oil. Within FACTLOG the focus is on LPG, which is formed as a result of distillation processes. An overview of the processes is shown in Figure 2. During these processes, some impurities (mainly pentane and sulphur in the form of hydrogen sulphur and mercaptan) that need to be removed using a complex set of interconnected processes. FACTLOG focuses on the route from LPG raw streams towards LPG refined streams. The main problem is how to achieve the proper quality of the final LPG streams, making sure the impurities are within legally set limits (chief among them being the sulphur content). The core idea is to detect possible trends and anomalies of the ingredient constitution in the early phases to minimise the impact in the final output tank.

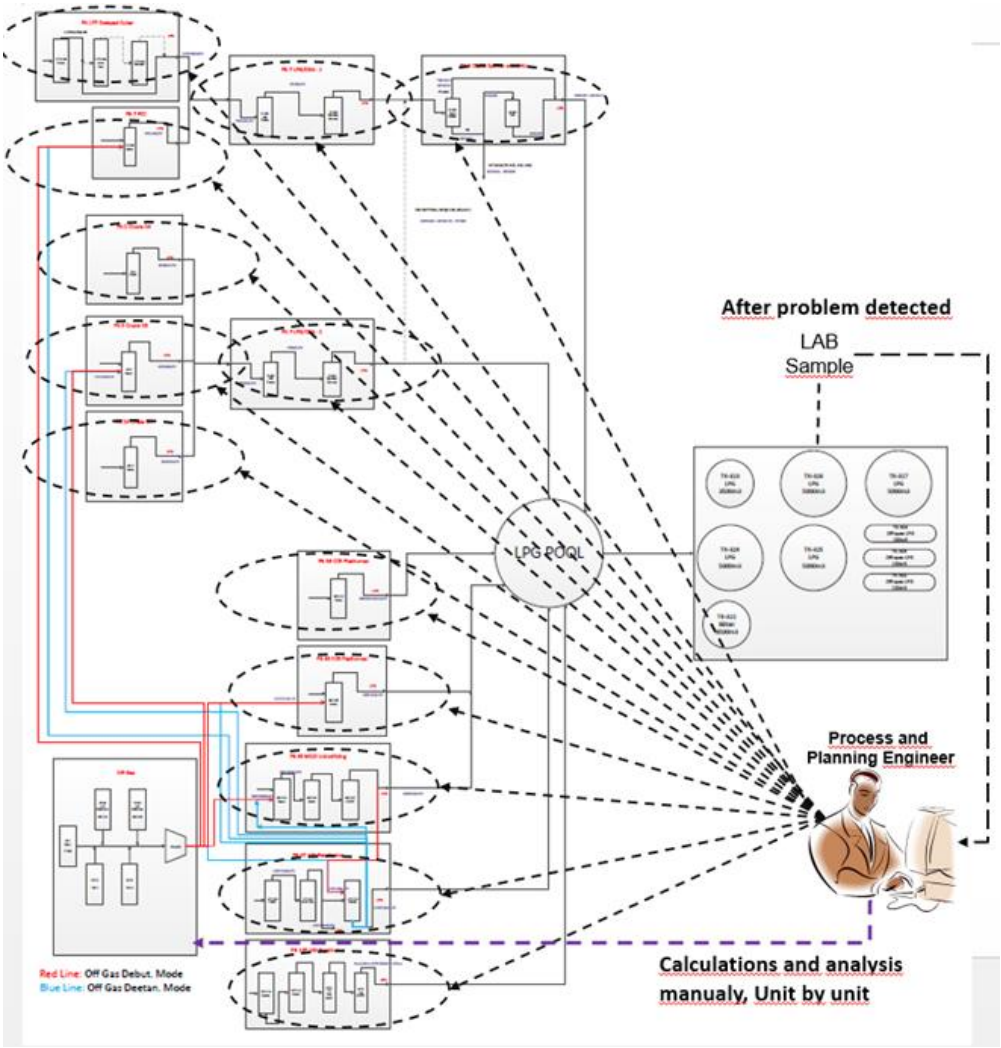

Figure 2: The Tupras LPG production process and the quality observation loop

2.2.1 Problem Scenarios and Requirements

There are three scenarios in the Tupras pilot.

Scenario #1: New Data Streams Monitoring and Storing

Description: LPG is the output of several different subsystems in the oil refinery, which means the system needs to collect, store and process a wide variety of data. While this is primarily a knowledge management problem where data needs to be transformed into a common format and validated, analytics can have a role in the cleaning and pre-processing.

Analytics approach: This scenario covers the very initial stages of analytics work, including data cleaning and preparation. The key requirement is to provide quality input for later processing. This includes data resampling and interpolation when sensors fail to report readings with reliable frequency. Some feature selection could also be performed at this point, using standard machine learning feature evaluation metrics to determine which data streams to focus on.

Scenario #2: Anomaly detection

Description: As the LPG production process is running, LPG is produced in several different sub-processes in subsystems of the refinery and is then collected into a common output pool (tank). LPG quality is tested on samples from this final pool. Since this is a laboratory test which takes time it would be highly beneficial if bad trends and anomalies would be detected earlier in the pipeline (and in time).

Analytics approach: This is an anomaly detection problem which can be tackled with a general approach of building a generative model of the process from historic data and then comparing the sensor readings to the model predicted values. When significant discrepancies are detected, an alert is raised. The generative model can either be based on typical value distributions of the sensor values; a set of regression models for individual stages of the pipeline (e.g. neural networks); process models built by experts or a combination of all three. Since the allowed limit values are known for all the impurities, a classification model could be built for each, predicting for some time horizon ahead that the limit value is going to be exceeded.

Scenario #3: Impact assessment and optimized intervention (LPG quality)

Description: As already described above in the description of scenario #2, the final LPG product is a mix of several different subsystem outputs. Once an anomaly is detected or off-specs product is predicted, the best way of remedying the problem needs to be determined. This includes identifying the relevant parts of the process/refinery and estimating the impact of these parts. The optimal intervention can then be found and performed.

Analytics approach: This is a root-cause identification and an optimisation problem solved by combining different components of the cognitive factory framework. The generative model of the pipeline, introduced in the approach to scenario #2, can be used to explore the impact of interaction with different parts of the pipelines. By exploring the space of possible interactions, the most impactful and effective interactions can be identified. The strategy of exploration is in the domain of the optimisation component.

2.2.2 Data Types and Sources

The Tupras data comes from their oil refinery in Izmit, Turkey. The refinery is equipped with a wide array of sensors with different properties. Here we provide an overview over the main parts of the pipeline and the different types of sensors.

There are ten LPG raw streams. The origin of these streams stems from 6 different units that are labeled as: Crude Distillation Unit (CDU), Platformer, Maximum Quality Diesel (MQD), Hydrocracker (HYC), Fluid Catalytic Cracking (FCC), Delayed Coker Unit (DCU). The crude oil is fed into crude distillation units (CDUs), which have the following main elements (described in detail in D1.1):

- Debutanizer column: removes the heaviest components (C5 and above)
- Deethanizer column: removes the lighter components (C1 and C2)
- DEA/merox column: removes sulphur (hydrogen sulphur and mercaptans)
- LPG recovery: recovers leftover LPG from by-products

The main types of sensors/values for these elements are:

- Temperature
  - Description (nature of data): temperature of the related stream
  - Measurement device: Temperature Transmitter / Thermocouple
  - Frequency: 1 data point / sec
  - Unit: °C
  - Range: 20 - 200 °C
- Flow
  - Description (nature of data): flow of the related stream
  - Measurement device: Flowmeter
  - Frequency: 1 data point / sec
  - Unit: m^3/h
  - Range: 0 - 3000 m^3/h
- Pressure
  - Description (nature of data): pressure of the related stream/ specific location of the column
  - Measurement device: Pressure Transmitter
  - Frequency: 1 data point / sec
  - Unit: kg/cm^2
  - Range: 0 - 20 kg/cm^2
- Level
  - Description (nature of data): level of capacity reached
  - Measurement device: Level sensor
  - Frequency: 1 data point / sec
  - Unit: %
  - Range: 0 - 100%

The purified LPG is then collected in the collection tanks which are equipped with the same sensors but the typical values have different ranges:

- Temperature: 0 - 40 °C
- Flow: -500 - +500 m^3/h
- Pressure: 0 - 8.5 kg/cm^2
- Level: 0 - 20 m

The chemical composition of the LPG and the levels of various impurities are measured by online analysers at different points in the process and in a laboratory using a gas chromatograph at the collection tank. The ranges of these readings differ depending on what impurity they are testing for:

- Sulphur: 0 - 250 mg/kg
- Butane: 0 - 200 %(mol/mol)
- Ethane: 0 - 300 %(mol/mol)
- Diene: 0 - 6 %(mol/mol)

The frequency of lab tests differs between the units from weekly tests in the final tank to daily in sulphur related units but are typically not performed more than once per day at best.

At least two years of historic data is available at the frequency of 1 data point/sec (for non-laboratory values). All values are real-valued numbers which are well suited for machine learning algorithms. The volume of data is very large and should be sufficient for analyses.

In case the data volume proves to be so bit that it is hard to process, down-sampling can be performed to produce a smaller dataset which is still representative.

2.3 Textile Industry: Pilot Case by PIACENZA

Piacenza manufactures woollen fabrics and is the leader in their market segment. Their plant receives the wool already cleaned and spun into yarns from a supplier and they perform the weaving into fabric on their machine looms. The looms are massive machines which must be set up with appropriately warping the yarns – a process that can take considerable time. The looms then run the weaving process during which a yarn may break. In case of such breakage, the process must be stopped and the yarn mended, slowing down production.

As a supplier to a very dynamic and demanding market, Piacenza continuously struggles with meeting the demand for their products. Optimally planning the production orders to minimise delays due to loom setup and adapting the plans to incoming high-priority orders is key to their business. Besides that, due to the demand there is a constant push to set the weaving process on the looms to go as fast as possible, but that increases the likelihood of yarn breakages. Predicting a yarn is likely to break during operation so that it may be avoided would help the process.

2.3.1 Problem Scenarios and Requirements

There are two scenarios in the Piacenza pilot.

Scenario #1: New data streams and storing

Description: The Piacenza plant already has a data management and collection system that stores the data from the existing sensors, the Manufacturing Execution System (MES) data, the Enterprise Resource Planning (ERP) data and the production schedule. In order to support the improvements aimed for in the project, the collected data needs to be extended with new data sources, in particular with regards to the quality of input materials (e.g., yarn for weaving) and from inside sources, including incremental output (e.g., fabric quality) and performance data (e.g., machine speed). At the time of writing this document the Piacenza team is working actively to extend the suite of machine sensors, but their utility and relevance must be evaluated.

Analytics approach: Analytics can help evaluate and identify relevant data streams among the new ones. By using methodology for feature selection, the predictive and explanatory value of individual data streams for the target events can be estimated and those that prove to not be useful in the pilot, can be dropped for the full deployment.

Scenario #2: Anomaly detection and new production plan formulation

Description: The plant constantly needs to plan how to process the work orders to meet the demand. Since the loom setup can take a long time, this is a crucial factor for planning. Some loom settings are more similar among each other than others any it may be beneficial to plan them one after the other on the same machine. The two main challenges to plan effectively are newly incoming orders of high-priority and the breakages on the looms. Both disrupt the regular operation and require the plant to modify the plans on-the-fly.

Analytics approach: The key contribution of analytics is to provide a prediction for when a yarn is likely to break. This is an anomaly detection problem, which can again be solved with a general anomaly detection methodology or using a targeted stream classification model if there is labelled data available. Since in the Piacenza case, the target anomaly is known in advance (breaking of the yarn), a targeted classification seems a more likely approach. An additional challenge is identifying the root cause, so that the loom operating parameters can be modified appropriately. This is a cognitive task as it combines several components. The predictions may be probabilistic but should be reliable enough so that when they are fed to the optimisation component as input, good plans can be produced.

2.3.2 Data Types and Sources

As mentioned above, Piacenza already collects some operational data in their plant, but plans to expand the set with new sensors. The existing and planned features along with their meta-information are listed in Table 8 in Appendix III – Piacenza Data.

The historic data contains the past orders and related planning from MES and ERP for looms and some sensor readings for energy consumption and water. Energy consumption is an important feature. In weaving, for example, given that all other parameters are constant, an increase in energy use reveals a wearing of components which can lead to an expected stop of production. Currently the past energy consumption is available at aggregated level for the whole weaving department and there are ongoing efforts to obtain this data at the machine level by installing further sensors. All the data relevant for the anomaly detection problem is numeric and well-suited for processing with machine learning algorithms.

At current stage of the project (September 2020) all the above-mentioned data are collected except the ones related with machinery consumption: they are still to be selected. This delay is related with the closure of the company due to the COVID emergency and to the period of mandatory holidays and layoff of the employees caused by the lack of orders. These provisions involved all the division of the company to grant equal economic treatment. It is expected to be able to run the selection and installation of the sensors by the end of 2020. In order to avoid any slowdown in FACTLOG activities a contingency plan has been formed: the involved partners have agreed to define the format of the expected data related with energy in advance and, eventually, to work on estimated dummy data until the real ones will be available. The details will be described in the deliverables on the data collection framework (D6.1 and D6.2).

2.4 Automotive Manufacturing: Pilot Case by CONTINENTAL

Continental is among the top worldwide electronics manufacturers. Its products are manufactured in electronics plants such as the plant in Timisoara where the pilot line is located. This plant produces high electronic products designed by different Continental developers worldwide. The products are customized for the final customer (i.e. automotive original equipment manufacturers) from the design phase onwards. Although these products (e.g. airbag control units, chaises controllers, hand brake controllers etc.) have a high complexity degree, their manufacturing process can be described (in brief) as follows:

- **SMT (Surface Mount Technology) lines:** High automated lines where electronic components are placed on the PCB boards.
- **PCBA (Printed Circuit Board Area):** PCB area, where the electronics built in SMT will be separated into smaller parts (PCB's) and tested electrically (In Circuit Test).

Additional processes can also take place in this area like Press Fit, Handling, Flashing of Microcontrollers and Temperature functional tests.

- **FA (Final Assembly) and Test Area:** This is the step of production where the electronics are connected to the mechanical part and finally tested and labelled. The processes in this area connect the mechanical parts: Screwing, Press Fit, Gluing, Riveting, Snap In. The testing area consists of tests line Functional test of the product, Automatic Optical Inspection, Force monitoring for the snap in, air leakage test.
- **Packaging and delivery operation:** In this step of manufacturing, the products are packed in customer specific boxes and all the information needed by customer is linked to the unique number of each box.

2.4.1 Problem Scenarios and Requirements

There are three scenarios in the Continental pilot.

Scenario #1: Machine downtime caused by breakdown

Description: In the Timisoara Continental plant all production lines run non-stop 24/7. Every unplanned downtime affects the plant because it could cause the situation of not delivering the needed quantity in time to the customers. Foreseeing malfunctions in advance would greatly alleviate this problem and improve the Overall Equipment Efficiency (OEE) by increasing the availability and increased quality of the Final Assembly Line. An example of a process where such monitoring is needed is the screwing process:

- The process is implemented with state-of-the-art technology components (e.g. screwdriver, screwdriver controller, axes systems for positioning, PLC for controlling the station).
- HMI interface with operator (permits the operator to know the status of the machine and the step sequence of the process).
- Specific communication with MES system for traceability and monitor performance of the line and specific process parameters.
- Issue handling process is manual: the operator sees an issue in HMI, tries to correct it by interaction with the machine. And in case of no solution the machine is set in breakdown.

Analytics approach: This is again a predictive maintenance problem solvable through either general anomaly detection or targeted (stream) classification models. By modelling the mechanical assembly area of the Final Assembly Line from the sensor readings we predict the machine malfunctions and improve (OEE). Our expectation is the reduction of down time caused by breakdown due to the possibility to forecast issues and plan them in preventive maintenance.

Scenario #2: Machine maintenance cost in % of total operational cost

Description: There are two different types of maintenance done in the plant: Preventive and Corrective/Reactive. By doing this, we can have two types of costs correlated with the number of failures. In case of preventive maintenance, we can have high costs with a low number of failures, while on the other hand, if we do not make preventive maintenance and wait until we have a big number of failures, we will, definitely, have high costs too and also unplanned downtime, which also produces a lot of costs. Finding a balance between the two is crucial.

Analytics approach: The challenge is finding the right balance in terms of cost of preventing malfunction vs number of failures in the Final Assembly Line which is an optimisation problem (illustrated by the graph in Figure 3). Another possible optimisation parameter is the reduction of maintenance cost in terms of head count time used for Predictive, Preventive and Corrective maintenance in the Final Assembly Line. Analytics support this optimisation problem by providing a reliable estimate of the likelihood of malfunction from the models from scenario #1.

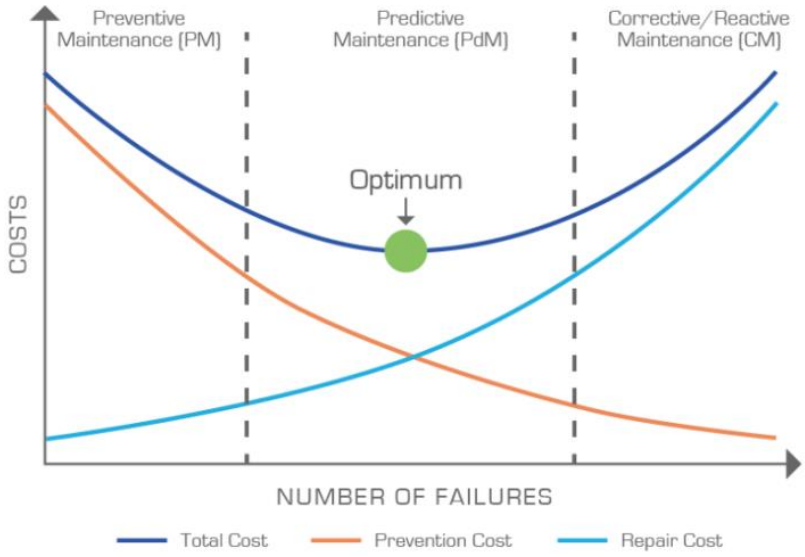

Figure 3: The optimisation of predictive vs. corrective maintenance cost

Scenario 3: Energy consumption of machines

Description: There is currently no correlation between the line ON/OFF state and the production scheduling at the plant. This way a lot of energy is consumed while the line is not running and the equipment is not operational/running/producing. By taking this into account at the schedule planning stage a significant amount of energy could be saved.

Analytics approach: This is an optimisation problem focused on the reduction of energy consumption/equipment by interacting with the planning in the production and also on a different implementation in the equipment used in production in terms of energy consumption control and monitoring. This includes:

- automatic setup of the Final Assembly Line status (plan / no plan) and Final Assembly Line energy consumption.
- automatic safe mode “shut down” and “wake up” based on production scheduling.

Analytics support this by modelling the energy consumption of machines using regressive machine learning models where this is not possible to estimate otherwise.

2.4.2 Data Types and Sources

The data in the Continental pilot is coming from the sensors in the production equipment. A Manufacturing Execution System (MES) which collects data from the machines is already in place. Each piece of equipment is connected to a MES client and they exchange data which is then stored in the MES database as illustrated in Figure 4.

Figure 4: The data collection process from the machines through MES clients

The format of the data and its other meta-information is presented in Table 9 located in Appendix IV – Continental Data. The sensors return binary or real-valued reading well suited for the kind of predictive and regression models needed by the pilot. Continental stores over ten years of historic data with a high resolution, so data volume may be a challenge. Feature selection and sub-sampling techniques will be used if necessary.

2.5 Steel Manufacturing: Pilot Case by BRC

BRC is the largest supplier of steel reinforcement in the UK, producing bespoke products for the construction industry. They use two types of raw materials, bars and coils of different diameters, in order to produce batches of final products made to customer specification. Each job requires a batch of steel to be produced in a certain shape – everything from simple straight bar to complex 3D shapes. The process involves cutting and shaping various diameters of steel reinforcing bar using various manual or automatic operations. The bending machines work with great forces and under great stress and do break down. Having a system that would supervise its operation and detect upcoming potential malfunctions would help avoid unnecessary down-time. Besides that, there is a need to improve management of the batches of steel in the production floor. An overview of the BRC production process is shown in Figure 5.

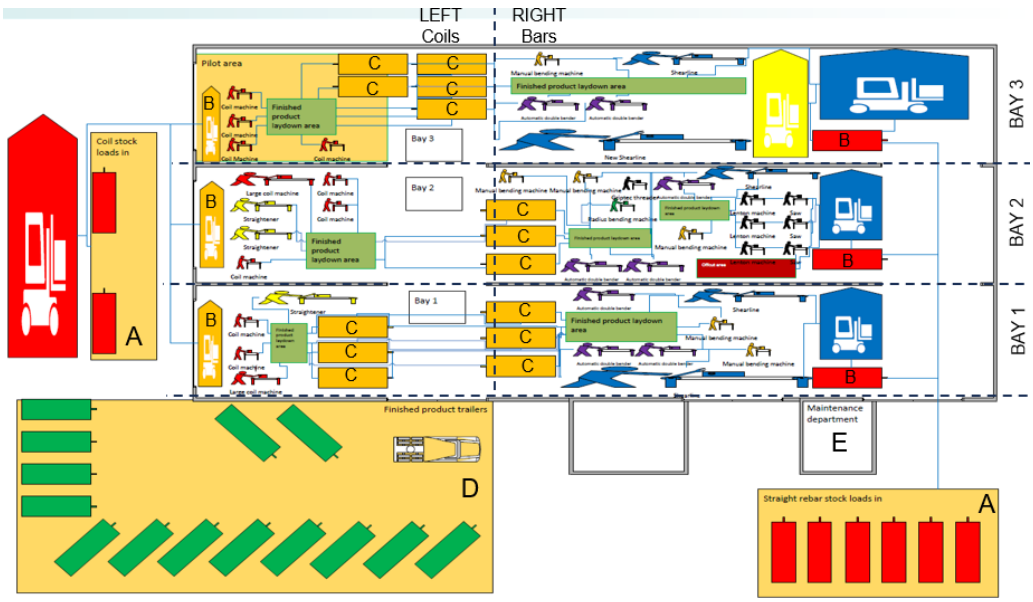

Figure 5: The BRC production process

2.5.1 Problem Scenarios and Requirements

There are two scenarios in the BRC pilot.

Scenario #1: Machine monitoring

Description: The hydraulic bending machine is one of the key pieces of manufacturing equipment at BRC. It works with immense pressure and is prone to breakage. The problem is, there is no way of foreseeing the upcoming mechanical failure until it occurs as the machine offers little insight into its operational status. The plan to overcome this is to install a sensory layer between the machine control panel and the machine itself to be able to capture the signals going to and from the machine. A system is needed that will predict the likelihood of breakage from the dynamics of these values.

Analytics approach: Similar as in the other cases this is an anomaly detection problem. Normal operational values need to be identified for different machine jobs and settings so that alerts may be raised when machine status deviates significantly. The challenge in this pilot lies in the fact that the machine sensor array is still being assembled and there is no existing data to learn from (an overview of the data, including the planned outputs of the new sensors, is presented in Appendix V – BRC Data). Therefore, more attention will need to be given whether to proceed with the more general approach of modelling the machine state with a generative model or perhaps to use more targeted stream classification models by incorporating more expert knowledge into the models. Nevertheless, this does not change the technical requirements of the pilot.

Scenario #2: Production scheduling and crane operation

Description: The steel rebar is processed in the BRC factory floor in batches. The batches are moved around using cranes when loading/unloading the materials to/from the machines. To achieve optimal plant operation, the jobs on the machines have to be planned together with the crane movements. This way, there is minimal waiting and new materials are provided to the machines when they finish previous jobs and the crane is available to move the processed batches away from the machines for storage or shipment.

Analytics approach: This is an optimisation problem where the space of possible plans needs to be searched to find the optimal one. The optimization component for the BRC case, depends upon parameters that are input data to be derived from analytics. As discussed in D1.1, the role of optimization in this pilot case is to provide solutions for BRC's complex multistage flowshop problem. In order for optimization to be able to derive to an optimal production schedule that takes under consideration raw materials, crane movement and machine maintenance, the analytics should provide indications with respect to, productions times, anomalies detection relevant to the machines' availability and schedules of maintenance. More precisely, the analytics should provide inputs with respect to operation and set up times for each product type in every production step. Such estimations could be easily derived for some products (for instance products with shape code C1-98) but much trickier for others (for instance products with shape code C99). Regression models predicting the operation times based on the machine state and product specifications (materials, shape…) can be used.

Additionally, when operating, different detected anomalies in the involved machines will have to be able to inform the optimization module for a potential problem (e.g.

underperformance based on currently produced batch). Lastly and in relation to the cranes detected anomalies in operation (e.g. availability, movement based on production schedule etc.) will also have to be identified in order to also inform the optimizer. These anomalies are outputs of the anomaly detection system described in scenario #1.

2.5.2 Data Types and Sources

Currently, the BRC production floor is not fully equipped for the application of digital twin technology for the scenarios described above. The process still needs to be equipped with sensors and some of it needs to be digitalised. These adaptations are being performed now in the scope of the project, including installing a sensory layer into the machine control board.

There are several different data sources in the BRC pilot:

- **Production data**: Production data that is stored is stored and utilised in the MES system however reports can be derived in excel
- **Planning data**: Planning data from the MES system using its existing optimisation based on machine providers production estimates and rules of optimisation we set. It is exported then to excel
- **Barmark data**: Data generated once entered from customer schedules into MES and exported by search query into excel
- **Transport data**: Mostly generated from the planning sheet however is then moved into a separate excel and the data is then moved in excel to generate loads
- **Machine capability**: Produced from experience and machine handbooks
- **Stock data**: Currently done by stock processor who records in excel
- **Machine sensor system**: Produced form new monitoring systems on the machine and data fed into a PLC then database
- **Crane sensor system**: Newly fitted system to measure distance of long travel and cross travel that's fed into a database
- **Scan data**: Dependant on existing MES System doing extra scanning and timestamping or if new app system for scanners needs to be created to work alongside
- **Machine PPM schedule**: Currently done manually and in existing Microsoft applications new system needs creating to work with FACTLOG system

The data parameters with their meta-information, including data types, data availability and data sources, are listed in Table 10 located in Appendix V – BRC Data. Note that though the majority of the values are numeric or binary, some of the parameters are less structured. A representative example of these is the parameter "Instructions for transport", collected from the planning sheet. Such values will need to be manually inspected by a data science expert to extract features appropriate for processing in machine learning algorithms. Most likely, manual transformations will be sufficient as there are only a few such values. Automatic feature generation approaches will be used if they prove necessary. For example, term frequency measures such as TF-IDF can be used to identify important phrases in the instruction texts and indicator features can be generated for those.

2.6 Overview

This section summarizes the preceding pilot-specific sections (sections 2.1 to 2.5) and gives an overview of the requirements for the analytics system – presented in Table 1. All the

pilots are, in one way or another, focused on detecting and preventing negative events in their production process. Since each such negative event is an anomaly with respect to "normal" production, this can always be framed as an anomaly detection problem. There is a distinction if a general anomaly detection approach is needed or a targeted approach can be used. The latter can be less resource consuming and might be more reliable but requires more specific learning data. The pilot-specific descriptions above point out where either of the approaches might be applicable. In Table 1 we point out the approaches which appear best at the moment of writing this document. Once we obtain more data from the pilots and gain more insights, some of the choices for optimal approach may change. Any changes in methodology will be reported in subsequent deliverables.

Table 1: Summary of analytics requirements

Pilot	Anomaly detection	Generative models / simulation	Stream classification / regression	Feature selection
JEMS	x	x	x	
TUPRAS	x	x		x
PIAC	x		x	x
CONT	x		x	
BRC	x		x	

The outputs of the analytics system are the inputs for the subsequent components which are triggered by the result of analytics or work on them. The optimisation services are a dependent component. Where the analytics have the job of detecting the problem, the job of optimisation is to find the actions that will rectify the situation in the best way. The relations between analytics and optimisation are already mentioned in specific pilots, here we summarise the main requirements from the optimisation perspective:

- **Indications with respect to anomaly detection** for the different units involved in the production, as a whole and per unit involved. The latter must examine and should have as a basic goal the identification/prediction/estimation of abnormalities in production. The sooner such a situation is identified, the sooner Optimization will be utilized to resolve the situation, hence the less energy (or, similarly, cost) will be required to recover to normal production.
- **Modelling through analytics and Machine Learning (ML) of the transformation process of process units that participates within production process.** Each process unit transforms input into output. For example, in the Tupras case a debutanizer receives a specific feed and applies temperature at the top and at the bottom as well as pressure in order to remove impurities, i.e., C1 and C2 from the top and C5, C6 etc. from the bottom. The different settings that may be applied (e.g., higher/lower temperature at the top/bottom with different levels of pressure applied) result in different outcomes with respect to the amount/percentage of impurities removed but they also correspond to different energy consumption/cost levels. Optimization requires incorporating all possible sets of settings for all related process units so as to select the ones for each process unit that collectively offer the best

trade-off for on-specs recovery between energy consumption/costs and improvement with respect to impurities. Modelling process transformation may be implemented through specific physical/chemical laws or it may be data-driven; i.e., process transformation models may be derived through regression or machine learning. In any case, Optimization assumes that for each process unit there exists a modelling of how it transforms input into output; analytics should provide such models for all process units where there exist data to do so.

- **Modelling specific values critical for the production.** Production processes can contain complex steps the properties of which are hard to fully anticipate. For example, in the BRC case it is hard to know how long the processing of a batch of steel will take on the bending machine with the set parameters. Since this is critical for efficient scheduling of jobs, a reliable model is needed that can estimate runtimes of the jobs using machine learning from the data.

3 Analytics System Design

This section presents the conceptual and methodological design of the analytics system (in section 3.1), including the relation to the other components of the FACTLOG platform, as well as the technical design (in section 3.2), including the high-level architecture and the list of tools and models identified to address the pilot requirements.

3.1 Conceptual Design

The analytics system provides fundamental functions that support the operation of the cognitive twins and its subsystems. Its services and models are the basic elements which are one of the building blocks of the cognitive processes. Their main role is to model components and processes from historical data when reliable or efficient models cannot be built based on theory and expert knowledge.

High-level cognitive functions for detecting variations and understanding their causes and impacts require reliable models of manufacturing machinery and equipment as well as processes in manufacturing operation. To detect anomalies, we first need to understand what normal operation is; to determine root-causes of errors and problems, we need to have a way of figuring out what contextual influences impact the operation of the observed system; and to run optimisation of the manufacturing processes we need a way of knowing how its stages and components will behave with different inputs.

The information needed by the cognitive functions such as those listed in the previous paragraph can be computed by using machine learning methodology. When historical data is available, a machine learning model can be built which predicts the target values. A wide array of modelling algorithms exists (for an overview see section 3.2), but at a high-level their operation can be summarised into two main actions:

- **Learning** – This is the initial step when historical learning data is input into the algorithm and a model is produced as the output. The data is comprised of a set of examples, each containing a set of context variables and the target variable that needs to be predicted. For example, the context variables can be the settings of a machine and the properties of a piece of raw material and the target variable the time the machine needs to process the piece at the given settings. In the case of streaming data, this step is periodically repeated to keep the model aligned with the current data.
- **Prediction** – This is the operational step after the model has been fit to the data in the learning step. The model is queried with a novel set of context variables, not necessarily seen before in the historical data, and the model predicts the target variable. Note that the predicted value can be categorical (e.g. a “good” or “defective” product) or numerical (e.g. the processing time will be 13 minutes). Some algorithms can also provide the estimated confidence for the result.

By linking together models for individual stages of a process such as a production line, we can build a larger model capable of holistic simulation the production process. In such a setup the outputs of the previous stage along with the settings of the current stage represent the input for the model of the current stage and the outputs of the model of the current stage are used as input for the model of the next stage. Let’s take for example the simplified version of the JEMS synthetic fuel plant (Figure 6). By building models of the three stages

based on historic data and chaining them together we can simulate the whole workflow and predict the properties of the fuel from the properties of the input feedstock and the settings of the individual stages.

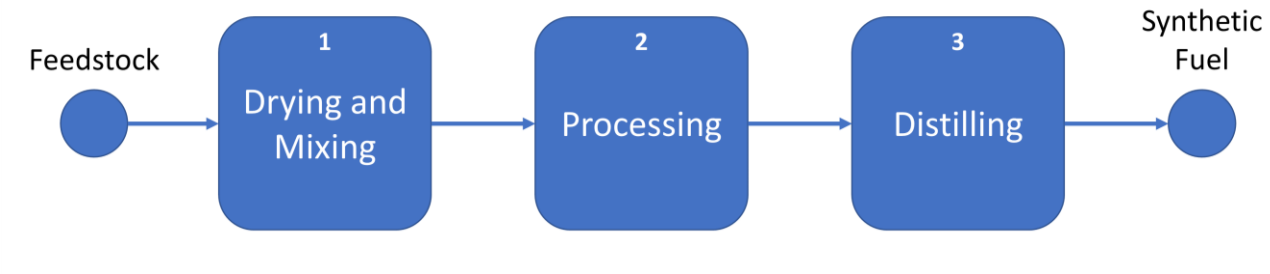

Figure 6: Simplified workflow in the JEMS synthetic fuel plant.

By following the Monte Carlo principle and running several such simulations by adding small perturbations in the inputs/outputs and settings to take into account possible errors and noise we can determine the most likely outcomes of the process. Having this prediction computed in advance, we can compare it to the sensor readings of the actual plant and raise an anomaly alert when they deviate significantly from the predicted values. By following the deviations back through the system, we can highlight the most likely root-causes of the anomaly.

Anomalies can also be detected by observing the historical data and identifying the typical states of the manufacturing system we are observing. By using clustering algorithms on the historical sensor values we can identify the typical states of the system and its transitions between them. When the system deviates from these states or transitions between them in an unlikely way an anomaly alert can be raised.

3.1.1 Relation to other FACTLOG components

The analytics system is well connected to other FACTLOG components. Some depend on its prediction outputs and some provide inputs for its algorithms. In this subsection the relations to the three main operational components (illustrated on Figure 7) are detailed.

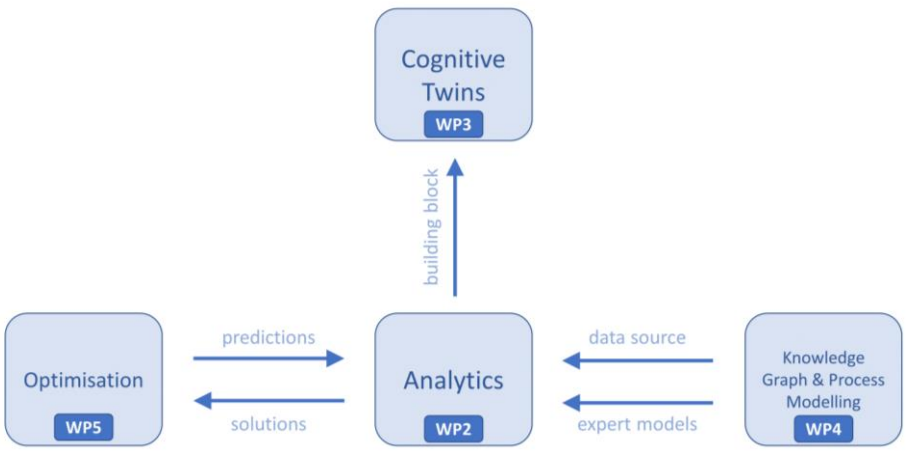

Figure 7: Relations between analytics and the three other operational FACTLOG component

Cognitive twins (WP3)

Analytics services are a building block of the cognitive twins. The full scale of the cognitive factory model and the explanation of its cognitive nature is beyond the scope of this document and will be described in detail in the deliverables from WP3. The analytics services play a role in the cognitive core of the cognitive framework (shown in Figure 8) – namely in reasoning, data cleaning and analysis and simulation and prediction services.

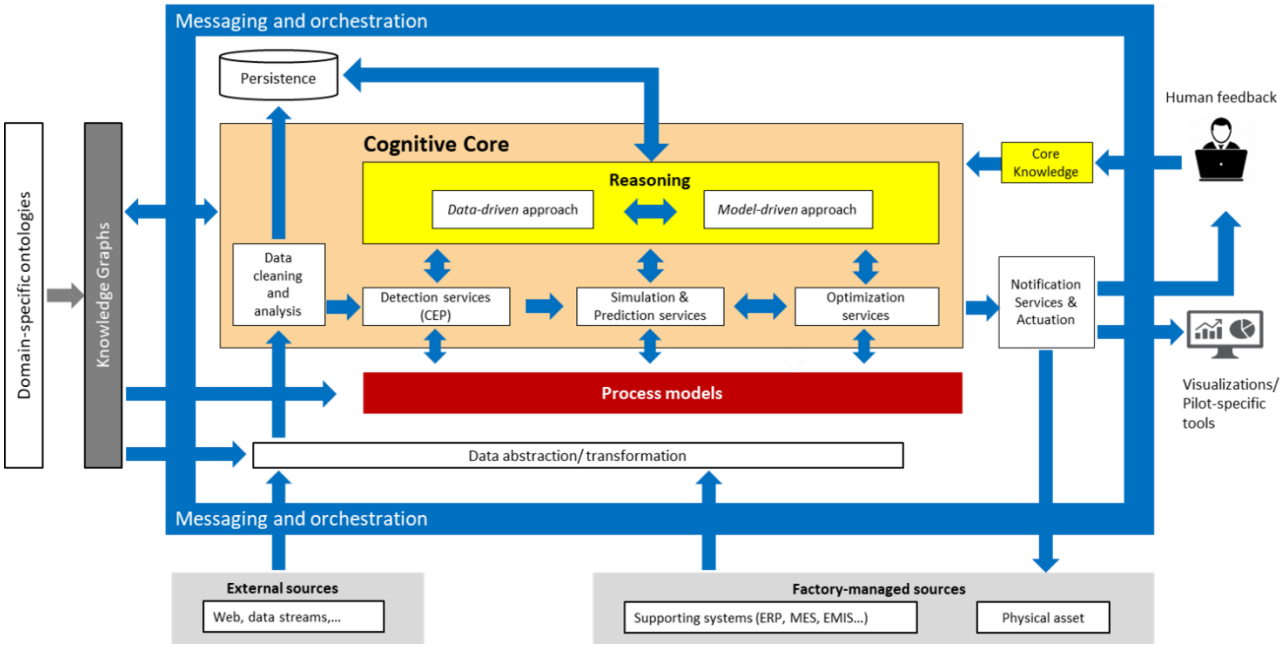

Figure 8: FACTLOG cognitive framework

In this document it is enough to take a more practically-oriented view in which the cognitive functions are higher-level and target specific operational goals. For example, running a predictive maintenance service for a factory needs analytics to process real-time sensor readings to identify an upcoming mechanical failure based on the historical experience. However, besides that it also needs the knowledge base to find out how the predicted failure impacts the shop floor and production plans and optimisation services to determine how to best plan the repairs needed.

Knowledge graphs and process modelling (WP4)

The data processed by the analytics services does not exist in a vacuum. Though the individual algorithms may in the end process a table of features, the preparation of this table along with the algorithm parameters depend strongly on the context, as does the interpretation of the results. This contextual data is stored in the knowledge graph which acts as the repository of data and settings. To put it directly, the knowledge graph is a data source for the analytics system. For example, it can answer questions such as:

- what is the acceptable error level of individual models;
- how often do we re-train the models for a system in operation whose state may be changing slowly with time;
- which of the sensors are relevant for the model?

The process models are similar to the analytics models in that they are representations of some of the same processes that are modelled in analytics based on past data. However, the process models are built using expert knowledge and theory. In many cases, the analytics models are used when such models cannot be built or are too costly. The analytics system can use them as priors for learning (for example in cases where the theoretical model may not sufficiently cover noise present in the operational system). The analytics models also need to be able to work together with the process models. For example, if we assume that when modelling the production line in Figure 6 we'd have a process model for stage 2 and analytics models for stages 1 and 3, we need to be able to chain them (i.e. use the outputs of one in another) to simulate the operation of the entire line.

Optimisation (WP5)

The optimisation algorithms search the configuration spaces of the domains where they run to determine the best configuration based on some criteria. In the context of manufacturing this can be the set of machine parameters or the order of jobs to perform on a production line or a combination of similar spaces. In this search they need reliable information on domain elements. For example, how long would a particular job take on a machine with the given input or how much material would a machine process per hour with the given settings. Analytics can provide some of these values based on historic data when they are otherwise not clear. Optimisation can query the analytics models during operation to ensure reliability and efficiency.

3.2 Technical Design

FACTLOG analytics system technical design is devised as a loosely coupled architecture. In order to avoid tight coupling, we make use of messaging queues and define a REST API interface. Both provide a uniform communication interface to underlying services and enable other users and services to consume them based on required resources or expected functionalities, abstracting them from the specific underlying architecture, services arrangement and infrastructure required to scale them.

Messaging queues enable to publish data ingested from data sources as well as data regarding state or computation results from any service, making it available to parties of interest that subscribe to the corresponding topics. Data of interest is also consumed by a persistence service, which stores it into a database so that can be later accessed for different purposes, such as analytics or eventually replay a series of events if needed.

This abstraction enables multiple services process the same data, even simultaneously if required, for different purposes. When doing so, streaming and batch processing can be applied, depending on use case requirements. This abstraction also enables proper decoupling from user interfaces, allowing not only to expose functionality through a web application, but also to build multiple tools, such as command line interfaces (CLI).

The architecture of the analytics system is shown in the diagram in Figure 9. As described above the different tools from the two partners implementing analytics in the pilots, JSI and NISSA, are all set up from the same configuration repository (envisioned to be the knowledge graph from WP4) and ingest data from the same persistent storage. Their outputs are communicated to other components, CEP services and any visualisation and pilot-specific tools through the messaging queue.

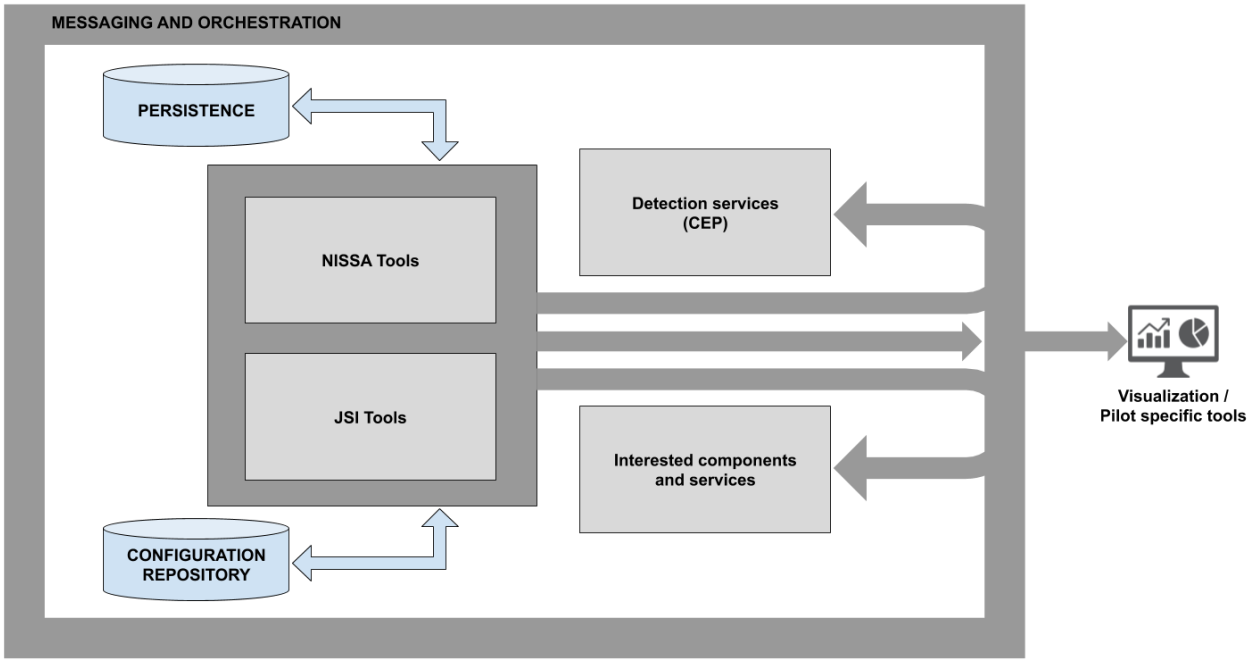

Figure 9: The analytics system architecture

In order to develop and maintain consistency across development and deployment environments, we make use of containerization technology, which allows to define isolated and predictable environments with software dependencies and runtime environment required by the application, that can be run anywhere. Containers provide an immutable definition of a service and its environment, so that multiple identical instances can be spawned without major effort. This has several benefits, such as versioning services with their runtime environment, and the reproducibility of issues across environments. In addition to this, we make use of standardized setup and runtime configurations for each service, persisted in JSON format, which are loaded into the containers and consumed by applications on start-up.

3.2.1 Technical tools

Below we describe some choices we made at technology level regarding tools, frameworks and platforms. FACTLOG services may use different sets of tools in order to achieve their purpose. The list provided in this section has been selected based on the pilot requirements described in sections 2.1 to 2.5.

QMiner

QMiner is an open-source analytics platform for real-time streams that may contain structured and unstructured data. It provides means for efficient storage, retrieval and analytics, while being able to respond in real-time. Written in C++, delivers strong performance to the users. QMiner was developed and is maintained at JSI and will serve as the basis for development of the analytics services. It contains fast and robust implementations of a wide variety of machine learning algorithms but can be further extended if needs arise.

scikit-multiflow

Is a machine learning library, focused on streaming algorithms, which allow to efficiently process streams, consuming limited memory and resources when doing so. Provides multiple algorithm implementations that enable incremental and adaptive learning. These algorithms incorporate means to detect changes in incoming data distribution, and thus can adapt how they learn. Among others, it provides implementations for the following algorithms: streaming linear regression, Hoeffding trees, Fast Incremental Model Trees with Drift Detections (FIMT-DD) and incremental versions of support vector machines (SVM) and neural networks.

ml-rapids

scikit-multiflow is one of the rare libraries that implements streaming algorithms in Python. According to our preliminary tests the library is, however, quite slow and includes inefficient implementations of the models. An in-house library, called ml-rapids, has been developed recently that implements some of the incremental learning models in C++ and exposes them via scikit-learn interface to Python or NodeJS.

scikit-learn

One of leading machine learning frameworks for batch learning algorithms. Provides implementations for data pre-processing, multiple algorithms such as logistic regression, SVM and random forest, and metrics to measure model's performance. Algorithms are implemented in Python, Cython, C and C++, ensuring great performance while still providing a great high-level interface to ease programming.

Probabilistic Soft Logic

Probabilistic Soft Logic is a statistical relational learning framework that provides means to model probabilistic and relational domains. Combines first-order logic and probabilistic graphical models (Markov random fields), being thus able to express complex phenomena as well as incompleteness and its derived uncertainty that we observe in the real world. By using soft logic, is able to reduce time complexity when computing results to a given query. In the context of FACTLOG, this framework is interesting as is can easily consume background knowledge from a knowledge graph and then perform global inference on incomplete data. In our specific case, we make use of the implementation provided by the University of Maryland and the University of California Santa Cruz.

Qlector LEAP

Qlector LEAP is a platform that provides solutions to many challenges faced by manufacturers, such as identify anomalies along production lines, predict organisational downtimes, provide demand forecasting and create schedules for planners and team coordinators in order to save time. This is achieved by using state-of-the-art artificial intelligence algorithms and by properly contextualizing obtained insights. A web user interface is provided to visualize KPIs, shop-floor layouts with metrics overlays, multiple analytics graphs as well as suggested actions that can be taken to mitigate detected issues. Access to different dashboards and platform functionalities is restricted based on users' profiles, considering their access level and manufacturing entities of interest. Mobile

notifications are issued to relevant actors, so that they can react on real time to issues identified in the platform.

3.2.2 Algorithms and models

In order to process the data, we make use of several algorithms and models from the fields of data analysis and machine learning. Below we describe a selection of those that we deem applicable to the FACTLOG pilots, making a distinction between batch and streaming algorithms.

Batch algorithms

- **Logistic regression**: models the probability of certain class computing the log-likelihood function on top of observations, considering they are independently Bernoulli distributed. Log-likelihood is maximized through gradient descent.
- **SVM**: is a non-probabilistic machine learning method that can be used for classification or regression. The algorithm constructs a hyperplane, where the vectors defining the hyperplanes are linear combinations with parameters of images of feature vectors. Points in the feature space are mapped to the hyperplane through the kernel function equalling a constant, property that can be used to understand how close test points from observed points in the train set are.
- **RNN**: are a specific architecture of neural networks, where connections between nodes allow to capture temporal dynamic behaviour. Can be used to predict future values, classify different time series as well as to produce a vector encoding of time sequences (auto-encoders).
- **LSTM**: are a specific architecture of neural networks, which contain feedback connections that enable processing sequences of data. LSTMs feature four components: a cell, an input gate, an output gate and a forget gate. Each cell has the responsibility of remembering values for arbitrary time intervals. The input gate receives new data, the forget gate is tasked with keeping memory of past states, while the output gate delivers some processing result.
- **Monte Carlo simulation**: are a family of algorithms that make use of random sampling that can be used to solve problems with a probabilistic interpretation, such as optimization or problems to be solved by generating draws from a probability distribution. Given a domain of possible inputs and a probability distribution of inputs, Monte Carlo algorithms require to generate random samples for inputs and perform some deterministic computation on them, to later aggregate results. Qlector LEAP implements this approach in multiple use cases. One of them is to provide accurate estimates and confidence intervals when estimating production plan termination dates.

Stream algorithms

- **Linear regression**: statistical algorithm that models a linear relationship between multiple explanatory variables and a scalar response.
- **Hoeffding trees:** are a specific implementation of classification and regression trees, which make use of the Hoeffding bound in order to decide if and when a node should be split. It assumes that the data distribution does not change over time.
- **FIMT-DD**: are a specific implementation of classification and regression trees that are able to learn from time-changing data streams, performing explicit change detection and the consequent adaptation.

- **Artificial neural networks**: models which consist of a set of simulated neurons, that can be used for classification or regression. The simulated neurons receive an input, have a set of weights which they adjust with every new training example using the backpropagation algorithm and then apply a non-linear transformation (activation function), before outputting to the following layer. Architectures have an input layer, one or more intermediate (hidden) layers and an output layer.

Appendix I – JEMS Data

Values returned by the sensors in the JEMS waste processing plant are described in the table below.

Table 2: JEMS data features

Name	Description	Value Type	Unit	Frequency
ES_104-01	Service switch status	BOOL		5
EA_104-01	Lime conveyor status	BOOL		5
XS_104-01	Start/Stop for Lime	BOOL		5
SC_104-01	Motor for climatization speed	BOOL	%	5
ES_103-01	Service switch status	BOOL		5
EA_103-01	Lime conveyor status	BOOL		5
XS_103-01	Start/Stop for catalyst	BOOL		5
SC_103-01	Motor for climatization speed	BOOL	%	5
LSH_500-02	High level in tank D500	BOOL		5
LSL_500-02	Low level in tank D500	BOOL		5
ES_500-01	Service switch status	BOOL		5
EA_500-01	Lime conveyor status	BOOL		5
XS_500-01	Start/Stop for Conveyor to P100	BOOL		5
SC_103-01	Motor for climatization speed	BOOL	%	5
TIC_100_12	Temperature in mixer vessel P100	REAL	°C	5
TIC_100_13	Temperature in evaporating column P100	REAL	°C	5
TIC_100_14	Temperature in evaporating column P100	REAL	°C	5
TIC_100_15	Temperature in evaporating column P100	REAL	°C	5
TIC_100_16	Temperature in evaporating column P100	REAL	°C	5
LSH_100-87	High level in tank P100	BOOL		5
LSL_100-84	Low level in tank P100	BOOL		5
LSM_100-85	Medium level in tank P100	BOOL		5
TSA_100-72	Thermostat in mixer P100 status	BOOL		5
pH_100-50	pH Measurement	REAL	pH	1
SC_100-21	Motor for mixer p100	BOOL	%	5
XS_100-21	Start/Stop mixer P100	BOOL		5
EA_100-21	Mixer P100 status	BOOL		5
YC_100-21	Torque measurement	REAL	Nm	
LC1_100-25	Load cell P100	REAL	kg	
LC2_100-26	Load cell P100	REAL	kg	
LC3_100-27	Load cell P100	REAL	kg	
ZSO_500-03	Knife between D500 and P100 - Opened	BOOL		5
ZSC_500-03	Knife between D500 and P100 - Closed	BOOL		5
ZSO_100-77	Valve back to P100 - Opened	BOOL		5
ZSC_100-77	Valve back to P100 - Closed	BOOL		5
ZSO_100-78	Valve from P100 to P120 - Opened	BOOL		5
ZSC_100-78	Valve from P100 to P120 - Closed	BOOL		5

TIC_100-80	Temperature on output of cooler	REAL	°C	5
TIC_100-81	Temperature on return to cooler	REAL	°C	5
ZI_120-54	Outlet valve from P100 to P120 - Position indicator	REAL	%	
ZSO_120-54	Outlet valve from P100 to P120 - Opened	BOOL		5
ZSC_120-54	Outlet valve from P100 to P120 - Closed	BOOL		5
XSD_120-54	Outlet valve from P100 to P120 - Direction	BOOL		
ZI_120-53	Outlet valve from P100 to P200 - Position indicator	REAL	%	
ZSO_120-53	Outlet valve from P100 to P200 - Opened	BOOL		5
ZSC_120-53	Outlet valve from P100 to P200 - Closed	BOOL		5
XSD_120-53	Outlet valve from P100 to P200 - Direction	BOOL		
SC_100-80	Motor for pump P100-80	BOOL	%	5
XS_100-80	Start/Stop for pump P100-80	BOOL		5
ZI_100-79	Outlet valve from P100 to P120 - Position indicator	REAL	%	
ZSO_100-79	Outlet valve from P100 to P120 - Opened	BOOL		5
ZSC_100-79	Outlet valve from P100 to P120 - Closed	BOOL		5
XSD_100-79	Outlet valve from P100 to P120 - Direction	BOOL		
ZI_100-52	Outlet valve from P200 to P120 - Position indicator	REAL	%	
ZSO_100-52	Outlet valve from P200 to P120 - Opened	BOOL		
ZSC_100-52	Outlet valve from P200 to P120 - Closed	BOOL		5
XSD_100-52	Outlet valve from P200 to P120 - Direction	BOOL		
SC_120-81	Motor for pump from P120 to P120	BOOL	%	5
XS_120-81	Start/Stop for pump from P120 to P120	BOOL		5
LC1_120-56	Load cell P120	REAL	kg	
LC2_120-57	Load cell P120	REAL	kg	
LC3_120-58	Load cell P120	REAL	kg	
TSA_120-72	Thermostat in mixer P120 status	BOOL		5
LSH_120-87	High level in tank P120	BOOL		5
LSL_120-84	Low level in tank P120	BOOL		5
LSM_120-85	Medium level in tank P120	BOOL		5
TIC_120_27	Temperature in vessel P120	REAL	°C	5
TIC_120_28	Temperature in evaporating column P120	REAL	°C	5
TIC_120_29	Temperature in evaporating column P120	REAL	°C	5
TIC_120_30	Temperature in evaporating column P120	REAL	°C	5
TIC_120_32	Temperature in evaporating column P120	REAL	°C	5
TIC_120_33	Temperature in evaporating column P120	REAL	°C	5
ZSO_120-77	Valve back to P100 - Opened	BOOL		5
ZSC_120-77	Valve back to P100 - Closed	BOOL		5
ZSO_120-78	Valve from P120 to P200 - Opened	BOOL		5
ZSC_120-78	Valve from P120 to P200 - Closed	BOOL		5
XSD_100-78	Valve from P120 to P200 - Direction	BOOL		
SC_120-80	Motor for pump P120-80	BOOL	%	5
XS_120-80	Start/Stop for pump P120-80	BOOL		5
LSH_200-87	High level in tank P200	BOOL		5
LSL_200-84	Low level in tank P200	BOOL		5

LSMH_200-86	Medium high level in tank P200	BOOL		5
LSML_200-85	Medium low level in tank P200	BOOL		5
LC1_200-15	Load cell P120	REAL	kg	
LC2_200-16	Load cell P120	REAL	kg	
LC3_200-17	Load cell P120	REAL	kg	
TIC_200_27	Temperature in mixer vessel 200	REAL	°C	5
TIC_200_28	Temperature in evaporating column P200	REAL	°C	5
TIC_200_29	Temperature in evaporating column P200	REAL	°C	5
TIC_200_32	Temperature in evaporating column P200	REAL	°C	5
TIC_200_33	Temperature in evaporating column P200	REAL	°C	5
TSA_120-72	Thermostat in mixer P200 status	BOOL		5
ES_200-75	Service switch status pump 200-75	BOOL		5
EA_200-75	Pump 200-75 status	BOOL		5
XS_200-75	Start/Stop for pump 200-75	BOOL		5
SC_200-75	Motor for pump 200-75	BOOL	%	5
PIC_200-29	Pressure in process tank P200	REAL	bar	5
ZSO_200-52	Outlet valve from P200 to turbine 106 - Opened	BOOL		5
ZSC_200-52	Outlet valve from P200 to turbine 106 - Closed	BOOL		5
XSD_200-52	Outlet valve from P200 to turbine 106 - Direction	BOOL		
ZI_200-52	Outlet valve from P200 to turbine 106 - Position indicator	REAL	%	
ZSO_200-14	Outlet valve from P200 to P400 - Opened	BOOL		5
ZSC_200-14	Outlet valve from P200 to P400 - Closed	BOOL		5
XSD_200-14	Outlet valve from P200 to P400 - Direction	BOOL		
ZI_200-14	Outlet valve from P200 to P400 - Position indicator	REAL	%	
ZSO_200-67	Outlet valve from P200 to turbine 108 - Opened	BOOL		5
ZSC_200-67	Outlet valve from P200 to turbine 108 - Closed	BOOL		5
XSD_200-67	Outlet valve from P200 to turbine 108 - Direction	BOOL		
ZI_200-67	Outlet valve from P200 to turbine 108 - Position indicator	REAL	%	
EA_106-10	Turbine 106 status	BOOL		5
XS_106-10	Start/Stop for Turbine 106	BOOL		5
SC_106-10	Motor for Turbine 106	BOOL	%	5
EA_108-10	Turbine 108 status	BOOL		5
XS_108-10	Start/Stop for Turbine 108	BOOL		5
SC_108-10	Motor for Turbine 108	BOOL	%	5
PIC_106-11	Pressure in turbine 106	REAL	bar	5
PIC_108-11	Pressure in turbine 108	REAL	bar	5
PIC_106-04	Pressure from turbine 106 back to P200	REAL	bar	5
PIC_108-04	Pressure from turbine 108 back to P200	REAL	bar	5
ZSO_106-12	Outlet valve from turbine 106 back to P200 - Opened	BOOL		5
ZSC_106-12	Outlet valve from turbine 106 back to P200 - Closed	BOOL		5
ZSO_108-12	Outlet valve from turbine 106 back to P200 - Opened	BOOL		5
ZSC_108-12	Outlet valve from turbine 106 back to P200 - Closed	BOOL		5
ZSO_120-51	Outlet valve from turbine 106 back to P100 - Opened	BOOL		5
ZSC_120-51	Outlet valve from turbine 106 back to P100 - Closed	BOOL		5

LC1_400-25	Load cell P400	REAL	kg	
LC2_400-26	Load cell P400	REAL	kg	
LC3_400-27	Load cell P400	REAL	kg	
TSA_400-72	Thermostat in mixer P400 status	BOOL		5
LSH_400-87	High level in tank P120	BOOL		5
LSL_400-84	Low level in tank P120	BOOL		5
PIC_400-29	Pressure in spare tank P400	REAL	bar	5
XS_120-81	Start/Stop for pump from and into spare tank P400	BOOL		5
SC_120-81	Motor for pump from and into spare tank P400	BOOL	%	5
ZSO_400-17	Outlet valve from P400 to P100 - Opened	BOOL		5
ZSC_400-17	Outlet valve from P400 to P100 - Closed	BOOL		5
XSD_400-17	Outlet valve from P400 to P100 - Direction	BOOL		
ZI_400-17	Outlet valve from P400 to P100 - Position indicator	REAL	%	
ZSO_400-15	Outlet valve from P400 to P100 - Opened	BOOL		5
ZSC_400-15	Outlet valve from P400 to P100 - Closed	BOOL		5
XSD_400-15	Outlet valve from P400 to P100 - Direction	BOOL		
ZI_400-15	Outlet valve from P400 to P100 - Position indicator	REAL	%	
ZSO_400-21	Outlet valve from P400 to P100 - Opened	BOOL		5
ZSC_400-21	Outlet valve from P400 to P100 - Closed	BOOL		5
XSD_400-21	Outlet valve from P400 to P100 - Direction	BOOL		
ZI_400-21	Outlet valve from P400 to P100 - Position indicator	REAL	%	
ES_400-03	Service switch status pump 400-03	BOOL		5
EA_400-03	Pump 400-03 status	BOOL		5
XS_400-03	Start/Stop for pump 400-03	BOOL		5
SC_400-03	Motor for pump 400-03	BOOL	%	5
ZSO_206-01	Valve to P300 - Opened	BOOL		5
ZSC_206-01	Valve to P300 - Closed	BOOL		5
XSD_206-01	Valve to P300 - Direction	BOOL		
ZI_206-01	Valve to P300 - Position indicator	REAL	%	
LSH_300-87	High level in tank P300	BOOL		5
LSL_300-84	Low level in tank P300	BOOL		5
LSMH_300-86	Medium high level in tank P300	BOOL		5
LSML_300-85	Medium low level in tank P300	BOOL		5
LC1_300-56	Load cell P300	REAL	kg	
LC2_300-57	Load cell P300	REAL	kg	
LC3_300-58	Load cell P300	REAL	kg	
TSA_300-80	Thermostat in mixer P300 status	BOOL		5
PIC_300-29	Pressure in second destilation tank P300	REAL	bar	5
TIC_300_27	Temperature in vessel P300	REAL	°C	5
TIC_300_91	Temperature in vessel P300	REAL	°C	5
TIC_300_28	Temperature in evaporating column P300	REAL	°C	5
TIC_300_29	Temperature in evaporating column P300	REAL	°C	5
TIC_300_30	Temperature in evaporating column P300	REAL	°C	5
TIC_300_32	Temperature in evaporating column P300	REAL	°C	5

TIC_300_33	Temperature in evaporating column P300	REAL	°C	5
ZSO_300-72	Valve from P300 to P400 - Opened	BOOL		5
ZSC_300-72	Valve from P300 to P400 - Closed	BOOL		5
YC_300-21	Torque measurement	REAL	Nm	
EA_300-21	Mixer in P300 status	BOOL		5
XS_300-21	Start/Stop for mixer in 3400	BOOL		5
SC_300-21	Motor for mixer in P300	BOOL	%	5
LSH_301-22	High level in tank P301	BOOL		5
LSL_301-21	Low level in tank P301	BOOL		5
LC1_301-28	Load cell P301	REAL	kg	
LC2_301-29	Load cell P301	REAL	kg	
LC3_301-30	Load cell P301	REAL	kg	
FQIR	Diesel flow	REAL	l	5
PIC_301-24	Pressure before storage tank	REAL	bar	5
EA_301-20	Pump P301-20 status	BOOL		5
XS_301-20	Start/Stop for pump P301-20	BOOL		5
SC_301-20	Motor for pump P301-20	BOOL	%	5
PIC_301-23	Pressure before P301	REAL	bar	5
LC1_202-51	Load cell P202	REAL	kg	
LC2_202-52	Load cell P202	REAL	kg	
LC3_202-53	Load cell P202	REAL	kg	
LSH_202-22	High level in tank P202	BOOL		5
LSL_202-21	Low level in tank P202	BOOL		5
PIC_202-23	Pressure in raw diesel tank P202	REAL	bar	5
PIC_202-44	Pressure after P202-25	REAL	bar	5
EA_202-20	Pump from P202 status	BOOL		5
XS_202-20	Start/Stop for Pump from P202	BOOL		5
SC_202-20	Motor for pump from P202	BOOL	%	5
EA_202-25	Pump from P202 status, flushing	BOOL		5
XS_202-25	Start/Stop for Pump from P202, flushing	BOOL		5
SC_202-25	Motor for pump from P202, flushing	BOOL	%	5
XS_202-47	Start/Stop for Pump from P202, flushing	BOOL		5
SC_202-47	Motor for pump from P202, flushing	BOOL	%	5
LC1_201-27	Load cell P201	REAL	kg	
LC2_201-28	Load cell P201	REAL	kg	
LC3_201-29	Load cell P201	REAL	kg	
LSH_201-22	High level in tank P201	BOOL		5
LSL_201-21	Low level in tank P201	BOOL		5
PIC_201-23	Pressure in water tank P201	REAL	bar	5
PIC_202-44	Pressure after P201-20	REAL	bar	5
XS_201-20	Start/Stop for Pump from P201	BOOL		5
SC_201-20	Motor for pump from P201	BOOL	%	5
EA_201-20	Pump from P201 status	BOOL		5

Appendix II – Tupras Data

This appendix contains the meta-information about the Tupras dataset, the naming schema of their features and an example of the dataset.

The **process features** follow the naming schema in Table 3 below.

Table 3: The standard naming schema of the process tags which are used in the Tüpraş İzmit Refinery.

Plant Name	Data Source Type	Transmitter ID	Extension	
0	XIC*	000	.PV	Process Value
			.SV	Set Value of Controller (Yokogawa DCS)
			.MV	Output Value of the Controller (Yokogawa DCS)
			.SP	Set Value of Controller (Honeywell DCS)
			.OP	Output Value of the Controller (Honeywell DCS)

Plant Name	Data Source Type	Transmitter ID	Extension	
0	XI*	000	.PV	Process Value (same for all DCS types)

X = F: flow, P: pressure, T: temperature, L: level

The naming of the data features that belong to **LPG storage tank** data follows the naming schema described below:

TPHTK: means it is a tank tag

XXX: id of the tank (LPG tank numbers: 302, 303, 304, 321, 322, 323, 324, 325, 326, and 327)

_____: which value do you want to see

TPHTKXXX_____.PV			
TPHTK	302/303/304/* 321/322/323/ 324/325/326/327	FLOW*	.PV
		TEMPERATURE*	
		LEVEL*	
		MASS	

		PRESSURE	
		STANDARD_DENSITY	
		VOLUME_AVAILABLE_NET	

* 302/303/304 only have flow, temperature and level sensors.

The features coming from the **online analyzers** are named according to the following schema:

AI: Analyzer Indicator

XXXXX: Analyzer ID

Root.GA3.177/ Root.GA1.147: Process unit where analyzer located

Root.GA3.177AIXXXXX.PV.Value

Root.GA1.147AIXXXXX.PV.Value

The features coming from the **laboratory analyses** are named according to the following schema:

02XXXLPGXX

02: Meaning it is a lab tag

XXX: Unit Name

XX: the abbreviation of the tested product/species (i.e. P: propane)

The table below, shows LPG raleted lab tags of plant-5 (CDU Debut 2)

025LPG13	PLT 5-LPG-1-3 BUTADIEN	**025LPGM**	PLT 5-LPG-METAN
025LPGB1	PLT 5-LPG-BUTEN1	**025LPGNB**	PLT 5-LPG-NBUTAN
025LPGC2	PLT 5-LPG-C2 BUTEN	**025LPGNP**	PLT 5-LPG-NPENTAN
025LPGC21	PLT 5-LPG-C2	**025LPGP**	PLT 5-LPG-PROPAN
025LPGC51	PLT 5-LPG-C5	**025LPGPP**	PLT 5-LPG-PROPILEN
025LPGET	PLT 5-LPG-ETAN	**025LPGRVP1**	PLT 5-LPG-RVP
025LPGIB	PLT 5-LPG-ISOBUTAN	**025LPGT2**	PLT 5-LPG-T2 BUTEN
025LPGIBL	PLT 5-LPG-ISOBUTILEN	**025LPGIP**	PLT 5-LPG-ISOPENTAN

Examples of data structure for all the data types are given in Table 4, Table 5, Table 6 and Table 7.

Table 4: Process sensors data example

Process Sensors									
		CDU_1			CDU_2	CDU_3			MQD
		2FIC350.PV	2FIC350.SV	2FIC350.MV	5TI496.PV	25TIC46.PIDA.PV	25TIC46.PIDA.SP	25TIC46.PIDA.OP	63FI1059.PV
Tag name		2FIC350.PV	2FIC350.SV	2FIC350.MV	5TI496.PV	25TIC46.PIDA.PV	25TIC46.PIDA.SP	25TIC46.PIDA.OP	63FI1059.PV
Unit of Measurement		M3/D	M3/D	%	DEGC				M3/HR
Description		2C-5 SARJ	2C-5 SARJ	2C-5 SARJ	HAD BUHAR CIKIS	DEBUT. SARJ GIRIS SIC.	DEBUT. SARJ GIRIS SIC.	DEBUT. SARJ GIRIS SIC.	63G101 AB CIKIS DEBUT.'A SARJ
The time interval can be longer. Up to 2 years. The frequency of all the FI/FIC/PI/PIC/TI/TIC is data point/sec	TimeStamp	2FIC350.PV - Snapshot - Value	2FIC350.SV - Snapshot - Value	2FIC350.MV - Snapshot - Value	5TI496.PV - Snapshot - Value	25TIC46.PIDA.PV - Snapshot - Value	25TIC46.PIDA.SP - Snapshot - Value	25TIC46.PIDA.OP - Snapshot - Value	63FI1059.PV - Snapshot - Value
	28/07/2020 14:54	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data
	28/07/2020 14:55								
	28/07/2020 14:56								
	28/07/2020 14:57								
		Platformer_2			FCC	DCU			
		36PIC347.PIDA.PV	36PIC347.PIDA.SP	36PIC347.PIDA.OP	7PI504.PV	Root.GA3.177TI2714.PV.Value	Root.GA3.177FIC2608.PV.Value	Root.GA3.177FIC2608.SV.Value	Root.GA3.177FIC2608.MV.Value
Tag name		36PIC347.PIDA.PV	36PIC347.PIDA.SP	36PIC347.PIDA.OP	7PI504.PV	Root.GA3.177TI2714.PV.Value	Root.GA3.177FIC2608.PV.Value	Root.GA3.177FIC2608.SV.Value	Root.GA3.177FIC2608.MV.Value
Unit of Measurement					KG/CM2	C	Sm3/d	Sm3/d	%
Description		DEBUT TEPE BASINCI	DEBUT TEPE BASINCI	DEBUT TEPE BASINCI	7C-501 TEPE BASINCI	C203 TEPE SICAKLIK	G204AB CIKISAKIM	G204AB CIKISAKIM	G204AB CIKISAKIM
The time interval can be longer. Up to 2 years. The frequency of all the FI/FIC/PI/PIC/TI/TIC is data point/sec	TimeStamp	36PIC347.PIDA.PV - Snapshot - Value	36PIC347.PIDA.SP - Snapshot - Value	36PIC347.PIDA.OP - Snapshot - Value	7PI504.PV - Snapshot - Value	Root.GA3.177TI2714.PV.Value - Time Average - Value	Root.GA3.177FIC2608.PV.Value - Time Average - Value	Root.GA3.177FIC2608.SV.Value - Time Average - Value	Root.GA3.177FIC2608.MV.Value - Time Average - Value
	28/07/2020 14:54	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data
	28/07/2020 14:55								
	28/07/2020 14:56								
	28/07/2020 14:57								

Table 5: Lab analysis data example

Lab Analysis									
Tag name		025LPG13	025LPGB1	025LPGC2	025LPGC21	025LPGC51	025LPGET	025LPGIB	025LPGIBL
Unit of Measurement		%(V/V)	%(V/V)	%(V/V)	%(V/V)	%(V/V)	%(V/V)	%(V/V)	%(V/V)
Description		PLT 5-LPG-1-3 BUTADIEN	PLT 5-LPG-BUTEN1	PLT 5-LPG-C2 BUTEN	PLT 5-LPG-C2	PLT 5-LPG-C5	PLT 5-LPG-ETAN	PLT 5-LPG-ISOBUTAN	PLT 5-LPG-ISOBUTILEN
Value Type	TimeStamp	025LPG13 - Raw - Value	025LPGB1 - Raw - Value	025LPGC2 - Raw - Value	025LPGC21 - Raw - Value	025LPGC51 - Raw - Value	025LPGET - Raw - Value	025LPGIB - Raw - Value	025LPGIBL - Raw - Value
	17/05/2019 06:00	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data
	19/05/2019 06:00								
	21/05/2019 06:00								
	24/05/2019 06:00								
	28/05/2019 06:00								
	31/05/2019 06:00								
	02/06/2019 06:00								
Tag name		025LPGIP	025LPGM	025LPGNB	025LPGNP	025LPGP	025LPGPP	025LPGRVP1	025LPGT2
Unit of Measurement		%(V/V)	%(V/V)	%(V/V)	%(V/V)	%(V/V)	%(V/V)	KPA	%(V/V)
Description		PLT 5-LPG-ISOPENTAN	PLT 5-LPG-METAN	PLT 5-LPG-NBUTAN	PLT 5-LPG-NPENTAN	PLT 5-LPG-PROPAN	PLT 5-LPG-PROPILEN	PLT 5-LPG-RVP	PLT 5-LPG-T2 BUTEN
Value Type	TimeStamp	025LPGIP - Raw - Value	025LPGM - Raw - Value	025LPGNB - Raw - Value	025LPGNP - Raw - Value	025LPGP - Raw - Value	025LPGPP - Raw - Value	025LPGRVP1 - Raw - Value	025LPGT2 - Raw - Value
	17/05/2019 06:00	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data
	19/05/2019 06:00								
	21/05/2019 06:00								
	24/05/2019 06:00								
	28/05/2019 06:00								
	31/05/2019 06:00								
	02/06/2019 06:00								

Table 6: Online analyzer data example

Online Analyzer							
Tag name		Root.GA3.177AI2901A.PV.Value	Root.GA3.177AI2901B.PV.Value	Root.GA3.177AI2901C.PV.Value	Root.GA3.177AI2701A.PV.Value	Root.GA3.177AI2701B.PV.Value	Root.GA1.147AI1004A.PV.Value
Unit of Measurement		ppmv	%V	%V	%wt	ppmv	%
Description		LPG H2S ANALIZOR	LPG C2 ANALIZOR	LPG C5 ANALIZOR	FUEL GAZ H2 ANALIZORU	FUEL GAZ H2S ANALZORU	LPG C2 ANALIZORU
Value Type	TimeStamp	Root.GA3.177AI2901A.PV.Value - End - Value	Root.GA3.177AI2901B.PV.Value - End - Value	Root.GA3.177AI2901C.PV.Value - End - Value	Root.GA3.177AI2701A.PV.Value - End - Value	Root.GA3.177AI2701B.PV.Value - End - Value	Root.GA1.147AI1004A.PV.Value - End - Value
	09/08/2020 11:10	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data
	09/08/2020 11:11						
	09/08/2020 11:12						
	09/08/2020 11:13						
	09/08/2020 11:14						
	09/08/2020 11:15						
Tag name		Root.GA1.147AI1004B.PV.Value	Root.GA1.147AI1004C.PV.Value	Root.GA1.147AI1004D.PV.Value	Root.GA1.147AI1004E.PV.Value	Root.GA1.147AI1004F.PV.Value	
Unit of Measurement		%	%	%	%	ppmv	
Description		LPG C3 ANALIZORU	LPG IC4 ANALIZORU	LPG NC4 ANALIZORU	LPG C5 ANALIZORU	LPG H2S ANALIZORU	
Value Type	TimeStamp	Root.GA1.147AI1004B.PV.Value - End - Value	Root.GA1.147AI1004C.PV.Value - End - Value	Root.GA1.147AI1004D.PV.Value - End - Value	Root.GA1.147AI1004E.PV.Value - End - Value	Root.GA1.147AI1004F.PV.Value - End - Value	
	09/08/2020 11:10	numeric data	numeric data	numeric data	numeric data	numeric data	
	09/08/2020 11:11						
	09/08/2020 11:12						
	09/08/2020 11:13						
	09/08/2020 11:14						
	09/08/2020 11:15						

Table 7: Tank sensors data example

Tank Sensors							
Tag name		TPHTK325FLOW.PV	TPHTK325GTEMPERATURE.PV	TPHTK325LEVEL.PV	TPHTK325MASS.PV	TPHTK325MASS_AVAILABLE.PV	TPHTK325PRESSURE.PV
Unit of Measurement		M3/HR	DEGC	M	TON	TON	KG/CM2
Description		LPG	LPG	LPG	LPG	LPG	GAZ BASINC
Value Type	TimeStamp	TPHTK325FLOW.PV - Raw - Value	TPHTK325GTEMPERATURE.PV - Raw - Value	TPHTK325LEVEL.PV - Raw - Value	TPHTK325MASS.PV - Raw - Value	TPHTK325MASS_AVAILABLE.PV - Raw - Value	TPHTK325PRESSURE.PV - Raw - Value
	10/07/2020 17:50	numeric data	numeric data	numeric data	numeric data	numeric data	numeric data
	10/07/2020 17:53						
	11/07/2020 11:56						
	11/07/2020 11:59						
	11/07/2020 12:02						
	11/07/2020 12:05						
	11/07/2020 12:08						
Tag name		TPHTK325STANDARD_DENSITY.PV	TPHTK325TEMPERATURE.PV	TPHTK325TOTAL_NET_VOLUME.PV	TPHTK325TRANS_LEVEL_SETPOINT.S	TPHTK325VOLUME_AVAILABLE_NET.PV	
Unit of Measurement		KG/L	DEGC	M3	M	M3	
Description		LPG	LPG	LPG	LPG	LPG	
Value Type	TimeStamp	TPHTK325STANDARD_DENSITY.PV - Raw - Value	TPHTK325TEMPERATURE.PV - Raw - Value	TPHTK325TOTAL_NET_VOLUME.PV - Raw - Value	TPHTK325TRANS_LEVEL_SETPOINT. SP - Raw - Value	TPHTK325VOLUME_AVAILABLE_NET .PV - Raw - Value	
	10/07/2020 17:50	numeric data	numeric data	numeric data	numeric data	numeric data	
	10/07/2020 17:53						
	11/07/2020 11:56						
	11/07/2020 11:59						
	11/07/2020 12:02						
	11/07/2020 12:05						
	11/07/2020 12:08						

Appendix III – Piacenza Data

This appendix contains the meta-information about the Piacenza dataset.

Table 8: Description of the Piacenza data parameters

DESCRIPTION	DATA PROVIDER	VALUE TYPE	HISTOR. DATA	MEASUREMENT DEVICE
Planning data weaving department	ERP/MES	Schedule		/
Planning data finishing department	ERP/MES	Schedule		/
Energy consumption for looms	Not available	Numerical	Not available	Energy Meter (TBD)
Theoretical energy consumption for looms	Values derived from formula	Numerical	Not available	/
Energy consumption for finishing machines	For a subset of machines for finishing RAMA1 - GAS (m3) + Energy (Kw/h) RAMA2 - GAS (m3)	Numerical	yes, 1 year	Machine energy meter
Absolute amount of energy consumption (per department; i.e. weaving/finishing)	Weaving: Energy (Kw/h) Finishing: GAS (m3) + Energy (Kw/h)	Numerical	yes, 1 year	Department Meter
Absolute amount of energy consumption (whole plant)	Energy (Kw/h) Finishing: GAS (m3) + Energy (Kw/h)	Numerical	yes, 1 year	Department Meters
HVAC consumptions (per department; i.e. weaving/finishing)	Only air conditioning for department.	Numerical	yes, 1 year	Meter
Water cubic meters	Only for finishing department	Numerical	yes, 1 year	Meter

Appendix IV – Continental Data

In this appendix you can find the Data types and limits for Continental pilot line. The features meta-information is listed in Table 9. The columns are:

- **TestName** - Name of each single test done, or process parameter provided by the Equipment
- **Measured Value** - The value provided/measured by the Equipment
- **Result** - The result of the test/process step compared with the limits
- **LSL** - Lower limit of each test/process step
- **USL** - Upper limit of each test/process step
- **Format** - The type of the data (i.e. R6.2 – Real with max 6 digits before the coma and 2 digits after the coma)

Important note: There are also a lot of test steps from the Continental test equipment but are too many to be listed in this document. There are few thousands for each piece of test equipment depending on the products. All these types of data exist in the raw data provided by Continental.

Table 9: Features for the Continental dataset

TestName	Measured Value	Result	LSL	USL	Format
Force	5.2	P	-9999	9999	R6.2
Distance	55.3	P	-9999	9999	R6.2
Camera2	1	P	1	1	R6.2
Quantity	131	P	-9999	9999	R6.2
RPM	3200	P	-9999	9999	R6.2
Tension	100	P	-9999	9999	R6.2
Frequency	13	P	-9999	9999	R6.2
Intensity	10	P	-55	9999	R6.2
Height1	5	P	-9999	9999	R6.2
Height2	3	P	-9999	9999	R6.2
Height3	6	P	-9999	9999	R6.2
Height4	5	P	-9999	9999	R6.2
Height5	2	P	-9999	9999	R6.2
Height6	5	P	-9999	9999	R6.2
Height7	9	P	-9999	9999	R6.2
Torque1	132	P	-9999	9999	R6.2
Torque2	130	P	-9999	9999	R6.2
Torque3	102	P	-9999	9999	R6.2
Torque4	132	P	-9999	9999	R6.2
Torque5	131	P	-9999	9999	R6.2
Torque6	111	P	-9999	9999	R6.2
Torque7	92	P	-9999	9999	R6.2
Angle1	50	P	-9999	9999	R6.2
Angle2	56	P	-9999	9999	R6.2
Angle3	60	P	-9999	9999	R6.2

Angle4	80	P	-9999	9999	R6.2
Angle5	50	P	-9999	9999	R6.2
Angle6	55	P	-9999	9999	R6.2
Angle7	50	P	-9999	9999	R6.2
ScrewingTime	15	P	-9999	9999	R6.2
Force1	0	P	-9999	9999	R6.2
Force2	0	P	-9999	9999	R6.2
Capacity	6614	P	5440	8160	R8.0
Height1	1.71	P	1.2	2.2	R6.2
Height2	1.55	P	1.2	2.2	R6.2
Nest	5	P	-999	999	R6.2
NozzleTemp	216	P	200	220	R6.2
Camera	0	P	0	0	R8.2
GlueWeight	4.12	P	4	4.2	R6.2
DispSpeed	80	P	0	200	R6.2
Pin1X	-34.5	P	-34.9	-33.95	R6.2
Pin1Y	2.85	P	2.35	3.25	R6.2
Pin1Z	7.67	P	7.4	7.8	R6.2
Pin2X	-31.91	P	-32.4	-31.45	R6.2
Pin2Y	2.79	P	2.35	3.25	R6.2
Pin2Z	7.64	P	7.4	7.8	R6.2
Pin3X	-34.55	P	-34.9	-33.95	R6.2
Pin3Y	-2.88	P	-3.25	-2.35	R6.2
Pin3Z	7.59	P	7.4	7.8	R6.2
Pin4X	-31.96	P	-32.4	-31.45	R6.2
Pin4Y	-2.82	P	-3.25	-2.35	R6.2
Pin4Z	7.62	P	7.4	7.8	R6.2
Pin5X	-28.88	P	-29.3	-28.35	R6.2
Pin5Y	4.5	P	3.98	4.88	R6.2
Pin5Z	6.97	P	6.7	7.1	R6.2
Pin6X	-27.08	P	-27.5	-26.55	R6.2
Pin6Y	4.54	P	3.98	4.88	R6.2
Pin6Z	6.96	P	6.7	7.1	R6.2
Pin7X	-25.22	P	-25.7	-24.75	R6.2
Pin7Y	4.63	P	3.98	4.88	R6.2
Pin7Z	6.97	P	6.7	7.1	R6.2
Pin8X	99	P	-999	9999	R6.2
Pin8Y	99	P	-999	9999	R6.2
Pin8Z	99	P	-999	9999	R6.2
Pin9X	-21.58	P	-22.1	-21.15	R6.2
Pin9Y	4.55	P	3.98	4.88	R6.2
Pin9Z	6.95	P	6.7	7.1	R6.2
Pin10X	-19.73	P	-20.3	-19.35	R6.2
Pin10Y	4.52	P	3.98	4.88	R6.2
Pin10Z	6.95	P	6.7	7.1	R6.2

Appendix V – BRC Data

This appendix contains the meta-information about the BRC dataset.

Table 10: description of the BRC data parameters

Parameter	Description	Value type	Unit of Measure	Frequency	Data Source	Real-time, Past data or both
Complete Y/N	Whether a barmark is completed or not based on entry of barmark finish time	Y/N	Scan end time inputted from a production machine touchbox in the MES system	Every new barmark scan	Production data	Both
Loaded Y/N	Whether a barmark is loaded based on loading scan time	Y/N	A timestamp when entered into the MES system for when a load has been scanned onto a trailer	Every loaded barmark	Production data	Both
Barmark in process	Whether the barmark Is in process based on scan start information and ended when there is a finish scan	Y/N	Scan time for when a barmark has been started	Every barmark	Production data	Both
Customer jobsite No	The number that the system uses to identify a specific contract using a contract year and specific identity number	Numerical	N/A	Every new contract	Planning data	Both
Customer sequence No	The number used to schedule a set of barmarks to be produced for a jobsite	Numerical	N/A	Every new order	Planning data	Both
Delivery prep No (DP)	The number to identify the delivery schedule that the sequence will be loaded on	Numerical	N/A	Every delivery prep set-up	Planning data	Both
Delivery area	The area of the UK that the delivery shall go to	Qualitative	address (string)	N/A	Planning data	Both
Regional and district postcode	The district postcode for the area of delivery (this would usually be full postcode but is anonymised)	Numerical and alphabetical	N/A	N/A	Planning data	Both
Tonnage breakdown for sequence	The amount of tonnage for all diameters of bar on the sequence and a total tonnage for each then total for all	Numerical	Tonnes	N/A	Planning data	Both

Colour of barmark tag	Colour of tag that the barmark will have	Colour	N/A	N/a	Planning data	Both
Comments	Any comments for the delivery or specific packaging requirements	Qualitative	N/A	N/A	Planning data	Both
Delivery date	The date that delivery is due	Date	date	N/A	Planning data	Both
Barmark-Jobsite	The contract year and jobsite number for referencing	Numerical	N/a	When entered onto MES System	Barmark data	Both
Barmark- Seq	The sequence number reference	Numerical	N/a	When entered onto MES System	Barmark data	Both
Barmark- No	The number reference of the barmark	Numerical	N/A	When entered onto MES System	Barmark data	Both
Barmark-Grade of steel	The grade of steel that the barmark is produced out of	Numerical	N/A	When entered onto MES System	Barmark data	Both
Barmark - Quantity	The amount of bars required to be produced for the barmark	Numerical	# of	When entered onto MES System	Barmark data	Both
Barmark-Dimensions	The dimensions of the bar that is to be produced to if it is a shape to be bent	Numerical	mm	When entered onto MES System	Barmark data	Both
Barmark-Cutlength	The length of bar to be cut to produce the bar	Numerical	mm	When entered onto MES System	Barmark data	Both
Barmark-Diameter	The diameter that the bars are to be produced out of	Numerical	mm	When entered onto MES System	Barmark data	Both
Barmark-Shape	The BS8666 shape code reference for the bar to be produced	Numerical	code	When entered onto MES System	Barmark data	Both
Barmark-Fabrication weight	The weight the total number of bars will come to	Numerical	KG	When entered onto MES System	Barmark data	Both
Barmark-Number of threads	The number of threads a bar requires	Numerical	# of	When entered onto MES System	Barmark data	Both
Barmark-Couplers required	The couplers that are required for the bar if coupled	Numerical	Coupler code	When entered onto MES System	Barmark data	Both
Barmark-number of bends	The number of bends on the bars	Numerical	# of	When entered onto MES System	Barmark data	Both
Barmark-number of arcs	?	Numerical	# of	When entered onto MES System	Barmark data	Both
Organisation of loads by sequence	The transport sheet shall have jobsite and sequence reference built into a list of loads	Numerical	N/A	3 times a day for delivery setup changes	Transport data	Not in real-time currently

Number of drops	The number of drops on a trailer depicted by the different jobsite numbers in a block	Numerical	# of	3 times a day for delivery setup changes	Transport data	Not in real-time currently
Haulier	The haulier the trailer will be going with dependant on the number of haulier contracts being used	Qualitative reference	N/A	3 times a day for delivery setup changes	Transport data	Not in real-time currently
Trailer No	The trailer that the load is being loaded on to	Number	N/A	3 times a day for delivery setup changes	Transport data	Not in real-time currently
Vehicle tonnage	The amount of tonnage of the sequences split by bar category with a total in red	Number	Tonnage	3 times a day for delivery setup changes	Transport data	Not in real-time currently
Quantity	The quantity of barmarks in a sequence	Number	# of	3 times a day for delivery setup changes	Transport data	Not in real-time currently
Instructions for transport	Required packaging of the sequence	Qualitative	N/A	3 times a day for delivery setup changes	Transport data	Not in real-time currently
Delivery date	The required date of delivery	Date	Date	3 times a day for delivery setup changes	Transport data	Not in real-time currently
Delivery time	The required time for delivery	Time	Time	3 times a day for delivery setup changes	Transport data	Not in real-time currently
Transport	Any specific requirements of the trailer or haulier its going on	Qualitative/code	N/A	3 times a day for delivery setup changes	Transport data	Not in real-time currently
Production tag number	The number of the production tag which is related to the specific barmark for scan purposes	Number	N/A	Generated when a barmark is generated	Production data	both
Machine code	The machine ID code referencing the touchbox of that specific machine or group of machines for example Lenton	ID	Number	Machine ID produced when reports are pulled up	Production data	both
Shift	The shift that the bars are produced on	Qualitative	N/A	Produced when report is pulled	Production data	both
Depot code	The code of the depot on the systems	Code	N/A	Produced when report is pulled	Production data	both
Start scan time	The start time of production as entered by a scan of the production tag (barmark) done by the operator	Date and time	Date and time	Produced when report is pulled	Production data	both

End scan time	The end of production for the production tag (barmark) which will be entered once another tag is scanned into the system	Date and time	Date and time	Produced when report is pulled	Production data	both
Machine capability-Cutlength	The Qualitative for the machine to produce certain bar lengths	Numerical	mm	N/A	Machine capability excel	Not currently available
Machine capability-Bar Quantities	The amount of bars that can be put through a machine in a single cycle with bar diameter as the variable	Numerical	# of	N/A	Machine capability excel	Not currently available
Machine capability-bending segments	The amount of bends a machine can do in a single cycle	Numerical	# of	N/A	Machine capability excel	Not currently available
Machine capability-Horizontal bending	The Qualitative to do bending on the x-y axis and maximum dimensional measurement	Numerical	mm	N/A	Machine capability excel	Not currently available
Machine capability-Vertical bending	The Qualitative to do bending on the z axis in up or down direction	Numerical and Y/N	mm	N/A	Machine capability excel	Not currently available
Machine workflow assignment	The workflow of the machines dependant on shape code, diameter and cutlength	Qualitative	N/A	N/A	Machine capability excel	Not currently available
Produce out of bar/coil	Produce the specific barmark out of bar or coil dependant on shape code, diameter and cutlength	Qualitative	Bar/coil/dependant	N/A	Machine capability excel	Not currently available
Bar stock	The amount of bar stock available in the factory	Numerical	Tonnage (individual bar numbers can be worked out through equation)	Availability updated every day	Stock data	Not real-time
Coil stock	The amount of coil stock available in the factory	Numerical	Tonnage (individual coil numbers can be worked out through equation)	Availability updated every day	Stock data	Not real-time
Machine downtime	A measurement of the downtime for the machine according to the bolt-on system	Time	N/A	Currently measured by MES when a operator scans code in	Machine sensor system	Not currently available
Machine ID	The machine ID of the bolt-on	Numerical	integer/string	per batch	Machine sensor system	Not currently available

Batch code	The barmark, sequence and jobsite as taken from MES system, matched by timestamping	Numerical	String	When pulled from MES system and what is currently on machine at the time matched with timestamp	Machine sensor system	Not currently available
Cycle Start time	The cycle times of each of the bars	Timestamp	Time	Every bar produced	Machine sensor system	Not currently available
Mains frequency	Real number for the frequency of electrical input	Numerical	Hz	Every step in cycle	Machine sensor system	Not currently available
Mains voltages	The voltage of the power supply going into the machine	Numerical	Voltage	Every step in cycle	Machine sensor system	Not currently available
Mains current	The current of the power supply going into the machine	Numerical	Amps	Every step in cycle	Machine sensor system	Not currently available
Hydraulic Temp	The hydraulic temperature measurement	Numerical	°c	Every step in cycle	Machine sensor system	Not currently available
Hydraulic pressure	The hydraulic pressure measurement of the system	Numerical	Pa	Every step in cycle	Machine sensor system	Not currently available
Step time	The time taken between each of the steps in a cycle for example feed,cut, feed,bend, feed, bend, feed and cut	Time	ms	Every step in cycle	Machine sensor system	Not currently available
Step power	The amount of power input for each step taken	Numerical	kW	Every step in cycle	Machine sensor system	Not currently available
Step Feed length	The amount of bar fed in for each step	Numerical	mm	Every step in cycle	Machine sensor system	Not currently available
Step angle bent	The angle measurement bent in a step	Numerical	Rad/degrees	Every step in cycle	Machine sensor system	Not currently available
Machine highlight issue	Issue detected if for example a oil leek occurs and hydraulic pressure drops severely	Y/N	N/A	Issue highlighted by system	Machine sensor system	Not currently available
Roller wear	The measurement of the wear on machine infeed and straightening rollers	Numerical	mm	Every step in cycle	Machine sensor system	Not currently available
Crane position	The positional position of a crane in relation to x-y co-ordinates	Co-ordinate	mm	Every RFID tag passed or equating system	Crane sensor system	Not currently available
Crane load identification	The identification of a load the crane has moved based on production tag data	Number	N/A	Every scan on barmark placement	Scan data	Both

PPM Plan	The planned preventative maintenance plan that is manually input to show what and when work is performed on a machine	Qualitative/ date	N/A	Input form engineering department	Machine PPM schedule	Not real-time currently
Workstation events	Machine stoppage codes that are entered into the MES system in order to identify a workstation (machine) event	Code	N/A	Whenever a stoppage happens	Production data	Both
Workstation event times	The timing of the workstation events	Time	N/A	When a stoppage is scanned in time will be entered into MES system	Production data	Both