

# Chapter

## Big Data Analytics in the Manufacturing Sector: Guidelines and Lessons Learned through the CRF Case

Andreas Alexopoulos<sup>1</sup>, Yolanda Becerra<sup>2</sup>, Omer Boehm<sup>3</sup>, George Bravos<sup>4</sup>, Vasilis Chatzigiannakis<sup>4</sup>, Cesare Cugnasco<sup>2</sup>, Giorgos Demetriou<sup>5</sup>, Iliada Eleftheriou<sup>6</sup>, Spiros Fotis<sup>1</sup>, Gianmarco Genchi<sup>7</sup>, Sotiris Ioannidis<sup>8,9</sup>, Dusan Jakovetic<sup>10</sup>, Leonidas Kallipolitis<sup>1</sup>, Vlatka Katusic<sup>5</sup>, Evangelia Kavakli<sup>6</sup>, Despina Kopanaki<sup>7</sup>, Christoforos Leventis<sup>9</sup>, Miquel Martínez<sup>2</sup>, Julien Mascolo<sup>7</sup>, Nemanja Milosevic<sup>10</sup>, Enric Pere Pages Montanera<sup>11</sup>, Gerald Ristow<sup>12</sup>, Hernan Ruiz-Ocampo<sup>6</sup>, Rizos Sakellariou<sup>6</sup>, Raül Sirvent<sup>2</sup>, Srdjan Skrbic<sup>10</sup>, Ilias Spais<sup>1</sup>, Giuseppe Danilo Spennacchio<sup>7</sup>, Dusan Stamenkovic<sup>10</sup>, Giorgos Vasiliadis<sup>9</sup>, Michael Vinov<sup>3</sup>

<sup>1</sup>Aegis IT Research LTD, UK <sup>2</sup>Barcelona Supercomputing Center, Spain <sup>3</sup>IBM, Israel

<sup>4</sup>Information Technology for Market Leadership, Greece <sup>5</sup>Ecole des Ponts ParisTech, France

<sup>6</sup>University of Manchester, UK <sup>7</sup>Centro Ricerche FIAT, Italy <sup>8</sup>Technical University of Crete - School of Electrical and Computer Engineering, Greece <sup>9</sup>Foundation for Research and Technology, Hellas - Institute of Computer Science, Greece <sup>10</sup>University of Novi Sad - Faculty of Sciences, Serbia <sup>11</sup>ATOS, Spain <sup>12</sup>Software AG, Germany

**Abstract.** Manufacturing processes are highly complex. Production lines have several robots and digital tools, generating massive amounts of data. Unstructured, noisy and incomplete data have to be collected, aggregated, pre-processed and transformed into structured messages of a common, unified format in order to be analysed not only for the monitoring of the processes but also for increasing their robustness and efficiency. This chapter describes the solution, lessons learned and guidelines for Big Data analytics in two manufacturing scenarios defined by CRF, within the I-BiDaaS project, namely ‘Production Process of Aluminium die-casting’, and ‘Maintenance and monitoring of production assets’. First, it reports on the retrieval of useful data from real processes taking into consideration data privacy policies and on the definition of the corresponding technical and business KPIs. It then describes the solution in terms of architecture, data analytics, and visualizations and assess its impact with respect to the quality of the processes and products. For the case of aluminium die-casting, the solution allows end-users to timely check the

status of the process and classify the quality levels; for the maintenance and monitoring case, they can check outliers for continuous and periodic control of the service conditions with a structured foundational database.

**Keywords:** Big Data, Self-service solution, Manufacturing, Die-casting, Advanced analytics and visualizations

## 1 Introduction

The manufacturing industry transforms material or assembles components to produce finished goods that are ready to be sold in the marketplace. The organizational structure of manufacturing companies is very complex and involves many business and operative functions with different roles and responsibilities in order to guarantee efficiency at every level [1]. The fourth industrial revolution [2] has initiated many changes in the industrial value chain, transforming the shop floor, which is the production part of the manufacturing industries. Companies are introducing process equipment provided with several robots and digital tools. In this way, it is possible to set and control processes in an automated manner that speeds up production with a high level of accuracy. Furthermore, large volumes of data are generated every day that may be collected and analysed for increasing process robustness and efficiency and building a technical cycle that reduces the consumption of energy and material. However, despite the potential benefits offered by the exploitation of Big Data, its usage is still at an early stage in many manufacturing companies [3].

Centro Ricerche FIAT (CRF) is one of the main private research centres in Italy and represents Fiat Chrysler Automobiles (FCA) in European and national collaborative research projects. In the context of the European Horizon 2020 I-BiDaaS project [4], CRF identified two use cases, in which complex datasets are retrieved from real processes. By exploiting Big Data analytics in these two cases, CRF aims to improve the process and product quality in a much more agile way through the collaborative effort of self-organizing and cross functional teams, reducing costs due to further processing and predicting faults and unnecessary actions. This requires solutions that will allow manufacturing experts to interact with Big Data in order to understand how to easily utilize important information often hidden in raw data.

The aim of this Chapter is to demonstrate how advanced analytic tools can empower end-users in the manufacturing domain (see Section 5) to create a tangible value from the process data that they are producing, and to identify a number of best practices<sup>1</sup> for utilization of Big Data analytics in the manufacturing domain. Specifically, *the first-best practice (1) is the correlation between the value of Big Data*

---

<sup>1</sup> These best practices are clearly identified and highlighted throughout the chapter.

*technology and the skills of people involved in the data management process.* In other words, the solution developed in I-BiDaaS provides a self-service Big Data analytics platform that enables different CRF end-users to exploit Big Data in order to gain new insights assisting them to make the right decisions in a much more agile way.

The remaining of this Chapter is organized as follows. Section 2 describes the process followed for the identification of the Big Data requirements in the manufacturing sector and demonstrates how it was applied to elicit the requirements of the CRF use cases, which are imposed the design of the I-BiDaaS Big Data solution. Furthermore, CRF requirements guide the definition of the experiments for assessing the developed system, described in Section 3. The architecture of the I-BiDaaS solution is described in Section 4. Finally, Section 5 reports on the lessons learned, challenges and guidelines reflecting the experience of the I-BiDaaS project. Section 5 also provides the connection of the described work with the Big Data Value (BDV) reference model and its Strategic Research and Innovation Agenda (SRIA) [5]. Finally, Section 6 concludes the chapter.

## **2 Requirements for Big Data in the manufacturing sector**

Alignment between business strategy and Big Data solutions is a critical factor for achieving value through Big Data [6]. Manufacturers must understand how the adoption of Big Data technologies is related to their business objectives in order to identify the right datasets and increase the value of the analytics results. *Therefore tailoring Big Data requirements to the business needs is the second-best practice (2) reported in this Chapter.*

In more detail, the I-BiDaaS methodology for eliciting CRF requirements draws on work in the area of early Requirements Engineering (RE), which considers the interplay between business intentions and system functionality [7][8]. In particular, the requirements elicitation followed a (mostly) top-down approach whereby business goals reflecting the company's vision were progressively refined in order to identify the user requirements of specific stakeholder groups (i.e., data providers, Big Data capability providers and data consumers). Their analysis resulted in the definition of system functional and non-functional requirements, which describe the behaviour that a Big Data system (or a system component) should expose in order to realize the intentions of its users. This process was facilitated by the use of appropriate questionnaires. In the cases that information on the requirements was available (either collected in the context of the project setup phase, or identified through a review of related literature [9][10]) this was used to partly pre-fill the questionnaires and minimise end-users' effort. Evidently, users were asked to check pre-filled fields and ensure that documented information was valid and accurate.

Table 1 gives a summary of the CRF requirements. Although it provides only an excerpt of the elicited CRF requirements, it demonstrates the application of the I-BiDaaS way-of-working in the CRF use cases.

Table 1: CRF Big Data Requirements

	<b><i>Business Requirements</i></b>
R1	Improve and optimize business processes and operations (business goal)
R2	Improve monitoring and maintenance of production assets (business goal)
R3	Improve decisions about production line re-planning based on the analysis of maintenance data (business goal)
R4	Maintain efficiency (quality business goal)
R5	Cost reduction (KPI)
R6	Product / service quality (KPI)
	<b><i>User Requirements</i></b>
R7	Data is stored locally into the Manufacturing Execution System (Data provider requirement)
R8	Real-time data on the operating status of the machines is obtained from SCADA sensors in real time (Data provider requirement)
R9	MES and SCADA sensor data information will be combined and proceed to real-time re-planning (Big Data analytics provider requirement)
R10	Line Operators will only visualise the results (data consumer requirement)
R11	Data Scientist will customize and then analyse data (data consumer requirement)
R12	Process manager will collaborate with the data scientist to decide on the action to actuate as a consequence of the analysis (data consumer requirement)
	<b><i>System Requirements</i></b>
R13	The system should enable aggregation of both attribute level and transaction level data coming from a variety of internal data sources and in multiple formats (FR)
R14	The system should support multilevel access control at resource and application level (NFR)
R15	The system should enable near real-time re-planning (NFR)

In more detail, the strategic CRF business goal ‘*To improve and optimise business processes and operations*’ (R1), was refined into a number of more operational business goals that need to be satisfied through Big Data analytics (e.g. “R3. *To improve decisions about production line re-planning based on the analysis of maintenance data*”). In addition, a number of relevant KPIs (e.g., “R6. *Product / service quality*”) were defined that can be used to assess the proposed solution (see Section 3). Continuing, at the user requirements level, requirements were described in terms of the characteristics of different data sources that are planned to be used (requirements R7 and R8); the analytics capability of the proposed solution envisaged (R9); and the different interface requirements of the end-users that will consume the analytics results (R10 - R12). Finally, analysis of the above user requirements resulted in the generation of the system requirements, both functional (R13) and non-functional (R14 and R15). Although described in a linear fashion, the above activities were carried out in an iterative manner resulting in a stepwise refinement of the results being produced. The complete list of CRF requirements elicited is described in detail in [11].

Further to forming the baseline of the I-BiDaaS solution (see Section 4), these requirements also assist the definition of experiments as described in the following Section 3.

### 3 Use cases description and experiments' definition; technical and business KPIs

The aim of experimentation is to assist stakeholders' acceptance of any new Big Data solution. *The definition of appropriate experiments is thus another best practice (3) reported in this Chapter.* In particular, the definition of CRF experiments aims at evaluating and validating the I-BiDaaS solution and its implementation in the context of CRF use cases. It follows a goal-oriented approach, whereby the experiment's goal(s) towards which the measurement will be performed are defined; then a number of questions are formed aiming to characterize the achievement of each goal; and finally, a set of Key Performance Indicators (KPIs) and associated metrics is associated with every question in order to answer it in a measurable way. Such KPIs have already been defined at the business level during requirements elicitation (described in Section 2). However, they need to be further elaborated so that they can be mapped onto specific indicators at the Big Data solution level. This ensures that the experiments consider both business and technical requirements, whilst maintaining traceability among business and application performance. The definition of each experiment also involved the specification of the experiment's workload in terms of the use case datasets and type of analysis envisaged, as well as the definition of the experimental subjects that will be involved in the experiment, as reported in the following Sections 3.1 and 3.2.

#### 3.1 Production Process of Aluminium die-casting

The 'Production Process of Aluminium die-casting' use case generates complex datasets from the production process of the engine blocks. During the die-casting process, molten aluminium is injected into a die cavity, mounted in a machine, in which it solidifies quickly. In this case, we have a large number of interconnected process parameters that influence the flow behaviour of molten metal inside the die cavity, and consequently, the productivity and the quality. *The fourth-best practice (4) is to identify the type of data generated.* Data collected from several sources can be disorganized and in different formats and data may not be exploited.

In this use case, the data provided for the analyses consist of a collection of casting process parameters, such as piston speed in the first and second phase, intensification pressures and others. In addition to the process data, CRF also provided a large dataset of thermal images of the engine block casting process, under a hypothesis that there is a correlation among process data, thermal data and the outcome of the process.

For the mentioned complexity of the process, it is important to not only carefully design parameters and temperatures but also to control them because they have a direct impact on the quality of the casting.

An excerpt of the structure of the dataset is reported in the table below with some of the key parameters identified for the detection of the quality level:

Table 2: Excerpt of Production process of Aluminium die-casting real anonymised dataset

Data type classification	Source/level	Data Type
VA2	Process Parameter	Number
VM2	Process Parameter	Number
Sigma2	Process Parameter	Number
PM2	Process Parameter	Number
Result 1	Process Parameter	Number
Result 2	Process Parameter	Number

Analysis of the above dataset aims to predict whether an engine block will be produced correctly during the casting process in order to avoid further processing and scraps, which would lead to financial savings for the manufacturers.

To test the efficiency of the I-BiDaaS solution in this context, an experiment has been defined, as shown in Table 3. As seen in Table 3, the Business KPI ‘Product/service quality’ identified during requirements elicitation (see Section 2) was further elaborated in order to define appropriate metrics (quality control levels related to good and defective products) and to map it to appropriate indicators at the I-BiDaaS solution level (execution time, data quality, cost).

Table 3: Overview of the ‘Production process of aluminium die-casting’ experiment

<b>Experiment’s Goals</b>	To test the efficiency of I-BiDaaS solution in the context of correlating defects with the production process parameters.	
<b>Experiment’s Questions</b>	<p><i><b>Q1. What is the quality of the analytics results?</b></i>  Q1.1 What is the accuracy of new models with respect to internal CRF Aluminium Casting models?</p> <p><i><b>Q2. How efficient is the process of data analytics?</b></i>  Q2.1 How efficient is the performance of the analytics application (algorithm)?  Q2.2 How efficient is the visualisation of the analytics solution to allow a quick intervention with specific actions?</p>	
<b>KPIs</b>	<b>Indicator</b>	<b>Metric</b>
<b>Business Level</b>	Product quality	Quality control 1
		Quality control 2
		Quality control 3
<b>Application Level</b>	Execution Time	Time to produce automated decisions
	Data Quality	Accuracy of new models with respect to internal CRF Aluminium die-casting models
<b>Platform Level</b>	Cost	Cost regarding personnel time spent on using the system (for analysis process) E.g. time spent for data anonymization
<b>Experimental subjects</b>	Quality assurance and control managers, Data analysts, Financial administrators, Infrastructure engineers, IT security personnel	

For each KPI, a baseline value for evaluating the performance of the I-BiDaaS solution has also been defined. For example, an increase of 2-6 % of the quality control level related to good products and a decrease of 1-4% and 0.05-2 % of the two quality control levels related to defective products is sought in order to satisfy manufacturers' requests in terms of product quality.

### 3.2 Maintenance and Monitoring of production assets

For the 'Maintenance and monitoring of production assets' use case, data have been retrieved from sensors mounted on several machines (e.g., linear stages, robots, elevators) along the production line of vehicles. We focused on welding lines in which robots are used to assemble vehicles' components and flexibility is required for the continual changes of the types of components and vehicles. A data server gathers sensor data, which categorised into two different datasets, namely SCADA and MES. The SCADA dataset contains production, process and control parameters of the daily vehicle production and is structured as follows:

Table 4: Structure of the dataset for the SCADA data

	Id	Value	Unit	Timestamp
Example	667	49.75	Mg	23/04/2018

There are over 100 sensors and each one is identified by a specific number (id). The other columns report on the value of the specific sensor, the unit of measurement and the timestamp.

The MES dataset contains specific data associated with the type of vehicle being produced and is structured as follows:

Table 5: Structure of the dataset for the MES data

	Date	Time	OP020.Passo[20]	modello op 020
Format	Date	Hour	Boolean	Number
Example	06/10/2018	09:44:22	0	11

When OP020.Passo [20] changes from 0 to 1, a new vehicle enters into the area provided with sensors and modello\_op\_020 indicates the model of the vehicle being processed.

Analysis of this data aims to the prediction of unnecessary actions and the improvement of the efficiency of manufacturing plants by reducing production losses. Once again, an experiment has been defined in order to test the efficiency of the I-BiDaaS solution in this context. The key points of the 'Maintenance and monitoring of production assets' experiment are shown in Table 6. In particular, data was analysed to obtain thresholds for anomalous measurements for all sensors and to build a foundational database with the history of anomalies. In this way, I-BiDaaS provided a useful solution that can be used by manufacturers with all other information that they cannot or will not share.

Table 6: Overview of the ‘Maintenance and monitoring of production assets’ experiment

Experiment	Maintenance and monitoring of production assets	
Experiment’s Goals	To test efficiency of I-BiDaaS solution in the context of anticipation of maintenance events (alarm).	
Experiment’s Questions	<p><b>Q1. What is the quality of the analytics results?</b>  Q1.1 What is the accuracy of new models with respect to internal CRF models in use (geographical representation of the process)?</p> <p><b>Q2. How efficient is the process of data analytics?</b>  Q2.1 How efficient is the performance of the analytics application (algorithm)?  Q2.2 How efficient is the visualisation of the analytics solution to allow the workers a quick intervention with specific actions?</p>	
KPIs	Indicator	Metric
Business Level	Product/ Service quality	OEE JPH
	Cost reduction	Maintenance cost
Application Level	Execution Time	Time to produce automated decisions
	Data Quality	Accuracy of new models with respect to internal CRF models
Platform Level	Cost	Cost regarding personnel time spent on using the system (for analysis process) E.g. time spent for data anonymization
Experimental subjects	Quality assurance and control managers, Data analysts, Financial administrators, Infrastructure engineers, IT security personnel	

As shown in Table 6, the business KPIs reported during requirements elicitation were further elaborated to identify related metrics (Overall Equipment Effectiveness (OEE) and maintenance costs) and to map them on specific indicators at the Big Data solution level (execution time, data quality and cost). Furthermore, for each KPI, a baseline value for evaluating the performance of the I-BiDaaS solution has been defined. For example, the prediction of unnecessary actions and the improvement of the efficiency should reduce production losses and achieve greater competitiveness of the company by an increase of 0.05 % of the current Overall Equipment Effectiveness (OEE) and a decrease of 50 % in maintenance costs.

#### 4 I-BiDaaS solutions for the defined use cases

*The final best practice (5) reported in the following Sections relates to the development of a solution that satisfies Big Data requirements of specific use cases by mapping the identified functional and non-functional concerns into a concrete software architecture [12]. In particular, the general requirements reported in Section*



2 were further clarified, taking into consideration the specific context of each use case (described in Section 3), resulting in customised solutions per use case described in the following Sections 4.1 and 4.2.

For both use cases, data gathered from the production lines are sent to CRF, where they are manipulated and masked. After the anonymization, data are sent to the I-BiDaaS Platform, hosted in a Virtual Machine. This represents a bridge between the I-BiDaaS infrastructure and CRF internal server, created by the I-BiDaaS technical partners. The same bridge is used to send the analytics results to the production plant end-users, as seen in Figure 1.



Figure 1. Flow of data and results

#### 4.1 Production Process of Aluminium die-casting

In this Section, the architecture, data analytics, visualization and results for the ‘Production Process of Aluminium die-casting’ use case are described.

##### 4.1.1 Architecture

The architecture of this use case can be seen in Figure 2. The architecture consists of several well-defined components. The Universal Messaging component is used for communication with most of the other components. To start with describing the data flow for this use case, we first consider the dataset. Data is transferred from CRF’s internal server to the I-BiDaaS platform server. Therein, the data is pre-processed and cleaned – this step is important as the data needs to be prepared for model training and inference tasks. Then, the data is given to the Machine Learning algorithm from the I-BiDaaS pool of ML algorithms. In this use case, the model is a complex neural network implemented in PyTorch [13] and trained jointly from thermal images and sensor datasets. The Machine Learning component outputs two results: training metrics/results for visualization purposes – used in the Advanced Data Visualization component, and the trained model used for inference. Both these results are transferred through Universal Messaging. In the end, for inference purposes, the Model Serving (Inference) Service component is used. In the initial phases of development, before the real data is fully prepared (e.g., retrieved, anonymized, etc.), the architecture uses for an initial components development realistic synthetic data. The use of synthetic data can make the development significantly more agile, but is utilized with care and under a quality assurance process. (For example, a final trained ML model has to be delivered on real data). We refer to Subsection 4.1.5 for details on realistic synthetic data generation and quality assessment.

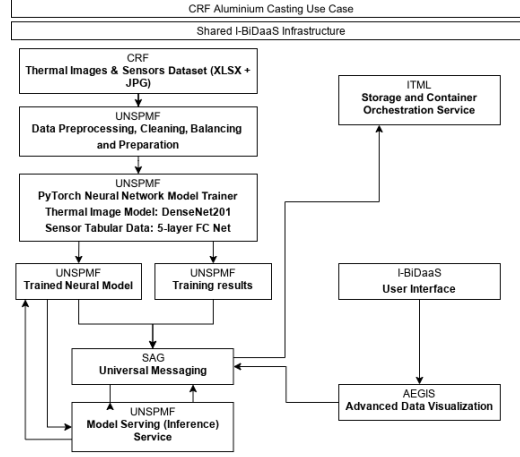


Figure 2. Architecture of the Production Process of Aluminium die-casting use case

#### 4.1.2 Data analytics

In this section, we describe in more detail the data analytics solution that corresponds to the four respective modules in Figure 2 (Data pre-processing, PyTorch neural network model, Trained model, and Training results) and that analyzes the thermal images and the sensors datasets. Under the hypothesis that there is a correlation among sensor data, thermal data and the outcome of the process, a further task is to classify combined image and sensor data inputs to see whether the cast engine blocks are without any production faults. Formally, data analytics here corresponds to an M-ary supervised classification task. As the dataset involves image classification, we utilize for this task Deep Convolutional Neural Networks [15].

The data is first pre-processed and normalized for training. Images were cropped and normalized and the sensor data was cleaned and normalized. First models of the experimentation protocol were shallow image classification models (without sensor data) but they were unable to perform well in this complex use case. Several modifications were tried out without measurable success: grayscale images, different hyperparameters, use of pretrained image models etc.

Subsequently, a joint model was developed where both engine thermal image data and engine sensor data were used. A deeper model was also used (DenseNet201[16][15]) and this model proved to be much better than the previously used shallow models and the ResNet18 [17] model. The DenseNet201 model was used in combination with a seven layer fully connected network for sensor data. The architecture of the model can be seen in Figure 3.

Another optimization was to use random under-sampling technique to balance the dataset, as the initial dataset was highly imbalanced with respect to the different classes. The dataset is understandably imbalanced as classes representing engine block faults are underrepresented. With these changes, the final model was created. The corresponding results are reported in Subsection 4.1.4.

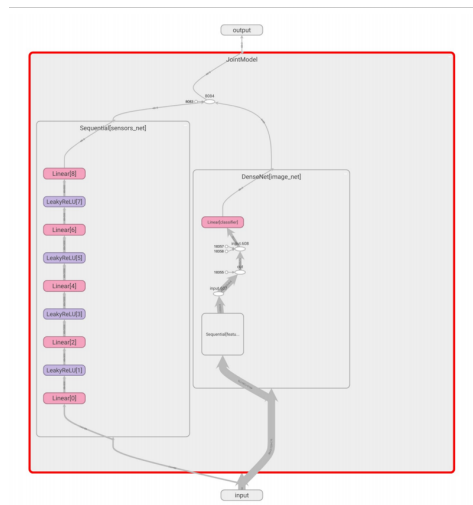


Figure 3. Architecture of the joint neural network

### 4.1.3 Visualizations

The approach to visualise the die-casting process results in real-time involves the deployment of a number of constantly updated visualisations which offer a complete overview of the results. These include the values of monitored sensor variables and the final classification of the end products of the process, as seen in Figure 4 and Figure 5.

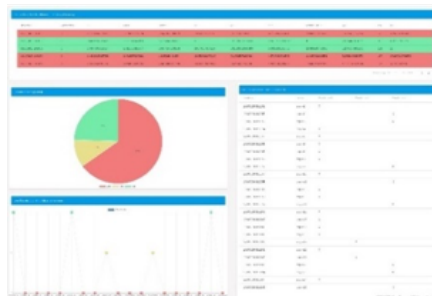


Figure 4. Real-time aggregated results



Figure 5. Real-time aggregated results

#### 4.1.4 Results

The models described in Subsection 4.1.2 were trained on both the original and the newly balanced datasets. Both models converge fully and quickly on the training datasets leading us to believe our architecture choice was correct. We favour the model trained on the balanced dataset as it learns to recognize faulty engine blocks

much better than the model trained on the imbalanced dataset, even though the overall accuracy is lower – simply because we have less faultless engines. In Figure 6, we see the accuracies of both models on the training and testing datasets (standard 80/20 split). The orange (top) line is the model trained on the full dataset and the pink line is the model trained on the balanced dataset [18][17].

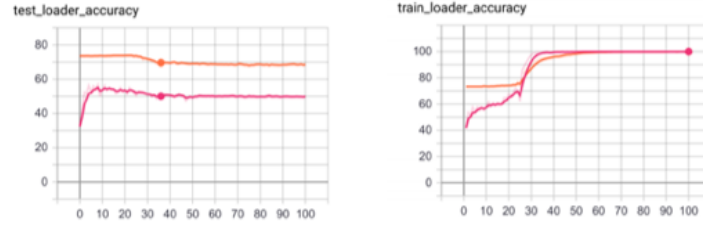


Figure 6. Training and testing accuracy for the two joint neural network models, when trained on full imbalanced data (orange line); and when trained on sub-sampled balanced data (pink line).

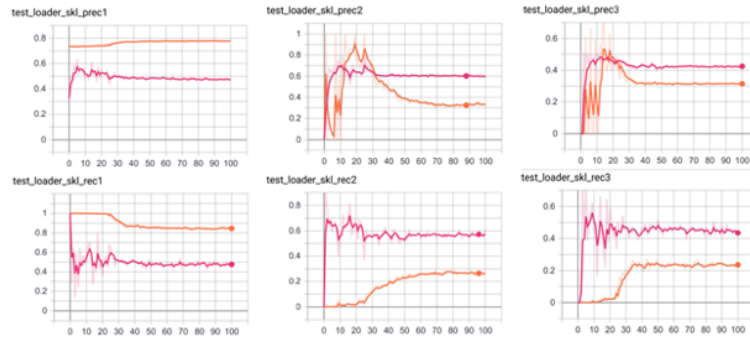


Figure 7. Precision and recall for different classes for the two joint neural network models for full imbalanced data (orange line) and sub-sampled balanced data (pink line)

When we investigate results per class, it is very clear that the balancing of the dataset is needed for the model to be able to recognize faulty samples better. This can be seen in Figure 7, especially in the case of class 2 and 3 in which are the most important classes for fault detection.

#### 4.1.5 Synthetic data generation and quality assessment

An initial development of the use case solution was carried out with realistic synthetic data. In parallel with the process of data anonymization, making data structured, etc., it was useful to carry out a synthetic data generation for early development stages with particular caution when extracting insights from synthetic data.

The fabrication of synthetic data that exhibits similar characteristics and similar distribution as the real data is a challenging task. The IBM Test Data Fabrication technology (TDF) was used for that purpose. TDF requires constraint rules that

model the relationships and dependencies between the data and leverages a Constraint Satisfaction Problems (CSP) solver to fabricate data that satisfies these constraints. The rules for the production of synthetic data were set by CRF with the help of IBM. The correlation between the real parameters and the synthesized parameters was further refined after reiteration of the data analysis.

For the initial evaluation of the synthetic data, we performed empirical and analytical validations. The empirical technique consisted of delivering these data to the expert production technicians, which were not able to indicate any difference with the actual production data, as there was no distinguishing factor for them. The second analytical technique was carried out by the CRF research team. They used the K-Means algorithm as their desired technique. Further evaluation was carried out by IBM while striving to perform a qualitative generic evaluation process for the real data compared with the fabricated data. This evaluation was concerned with methods to judge whether the distributions of the fabricated data and the original data were comparable, what is commonly referred to in the literature as the general utility of the datasets. In addition to the general utility, IBM also considered the specific utility, i.e., the similarity between the synthetic data and the original data.

The propensity mean-squared-error (pMSE) [19] was used as a general measure of data utility to the specific case of synthetic data. Propensity scores represent probabilities of group memberships. If the propensity scores are well modelled, this general measure should capture relationships among the data that methods such as the empirical Cumulative Distribution Function (CDF) may miss.

The method is a classification problem where the desired result is poor classification (50% error rate), giving better utility for low values of the pMSE.

Randomly sampling 5000 datapoints from the real and synthetic datasets, and using a logistic regression to provide the probability for the label classification, we were able to show that the measured mean pMSE score for the ‘Production Process of Aluminium die-casting’ dataset is 0.218 with a standard deviation of 0.0017.

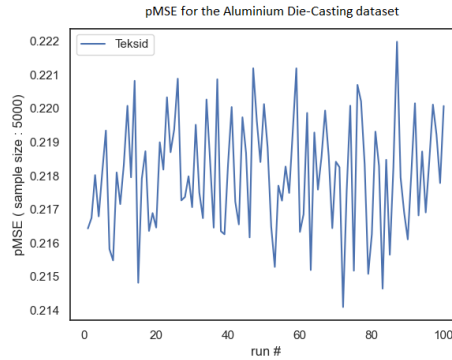


Figure 8. Results for 100 random sampling taken from the real and the synthetic data (5K datapoints each) and the pMSE calculated using a logistic model

## 4.2 Maintenance and Monitoring of production assets

In this Section, the architecture, data analytics, visualization and results for the ‘Maintenance and Monitoring of production assets’ use case are described.

### 4.2.1 Architecture

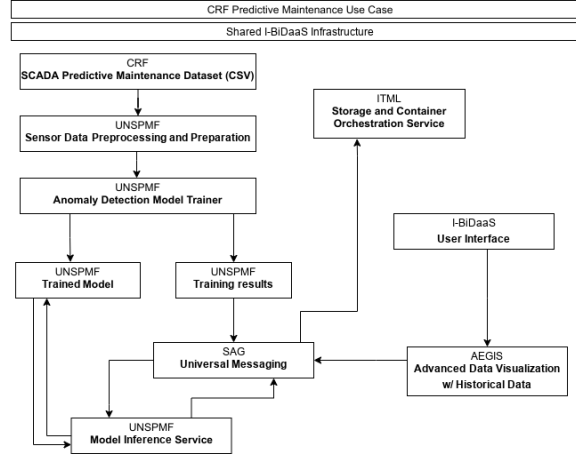


Figure 9. Architecture of the Maintenance and Monitoring of production assets use case

Figure 9 shows the architecture, which consists of several well-defined components. The Universal Messaging component is used for communication in most of the components. To start to describe the data flow, we start with the dataset. Data are sent from CRF to the I-BiDaaS platform. There, the data is pre-processed and prepared for model training with an outlier detection model. The outlier detection model outputs two results: training results for visualization purposes – used in the Advanced Data Visualization component, and the trained model used for inference. Training results are transferred through Universal Messaging. In the end, for inference purposes, the Model Inference Serving component is used. It is also important to say that all the components use containerized (i.e., Docker [14]) backbone from the Storage and Container Orchestration Service. Data is visualized and the jobs are scheduled through the I-BiDaaS User Interface component.

### 4.2.2 Data analytics

Data, described in Section 3, has been transformed into separate time series – one per sensor so that each sensor can be monitored separately. Since the measurements were not labelled (anomalous / non-anomalous), outlier detection algorithms arose as natural candidates for this use case [20]. We constructed an outlier detection model for each of the time series. While more advanced algorithms can be used, we adopted a simple, easy to implement, and computationally cheap, yet here effective solution, based on the Inter-Quartile Range (IQR) test. Results of these models

could be used for suggesting if a measurement is an outlier and for discovering the pairs of sensors that have anomalous measurements at similar timestamps. Preparation of these models was done using Python, and it consisted of the following steps:

1. For each sensor, obtain thresholds for anomalous measurements using a modified interquartile range (IQR) test. Three different variants of IQR-like tests were performed:  
 $(Q_1, Q_3) \in \{(5^{th}, 95^{th}), (10^{th}, 90^{th}), (25^{th}, 75^{th})$  where  $Q_1$  and  $Q_3$  are the corresponding percentiles;
2. With obtained thresholds, filter the time series such that only anomalous measurements were kept, as shown in Figure 12;
3. Calculate the Dynamic Time Warping (DTW) [21] distance between outlier time series.
4. Rescale distances to  $[0, 1]$ ;
5. Group pairs of sensors by the distance into groups:  
 $[0, 0.1), [0.1, 0.2) \dots [0.9, 1]$ .

Time series with anomalous measurements obtained in step 2 enabled us to see the outlier trends for each sensor and to compare their behaviour. Comparison of anomalous trends was made using steps 3, 4 and 5. If the distance obtained in step 5 is small, it means that two sensors output anomalous measurements in a similar fashion. Therefore, if one of them fails, then the other sensor in the pair should also be inspected.

#### 4.2.3 Visualizations

Data stemming from the aforementioned analysis are presented using a multi-step approach that allows operators drill down to sensory data and detected anomalies in an intuitive and easy to use way. Starting from a given month, operators then select the category of sensors they wish to see and immediately have an overview of the ones having anomalies detected. Upon selection of a sensor, operators see the anomalies detected during the selected month and can furthermore select a specific day to see the actual values and therefore review the actual anomaly that was detected.

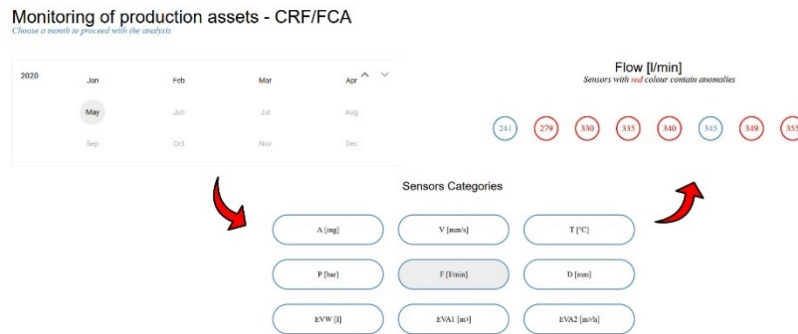


Figure 10: Sensor category selection

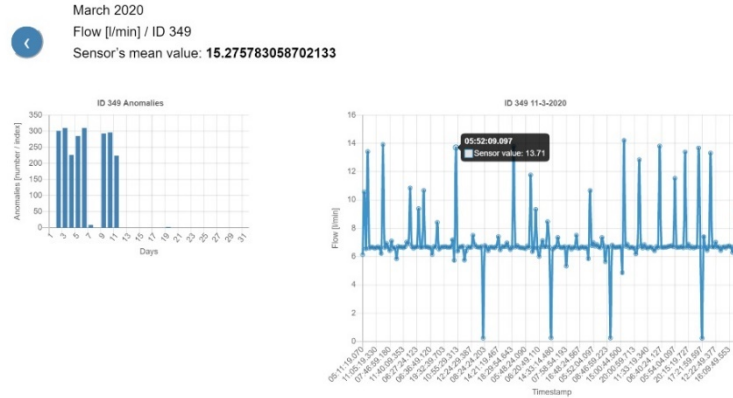


Figure 11: Sensor history and details

#### 4.2.4 Results

The obtained boundaries (from step 2 in Section 4.2.2) could be used for daily analysis of sensors and various visualization tasks, such as showing the number of anomalous measurements for the current day, comparing the number of outliers between two sensors for the given time window, etc. as seen in Figure 12 and Figure 13.

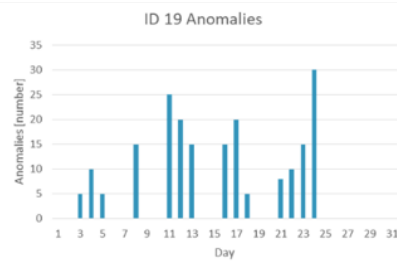


Figure 12: Number of anomalies per day

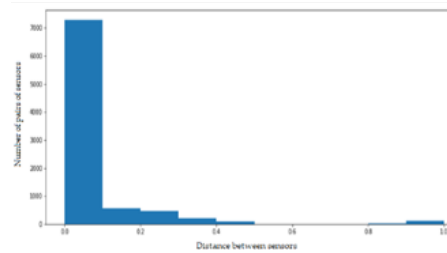


Figure 13: Number of outliers between sensors

## 5 Discussion

Reflecting on CRF's experience and best practices in the context of the I-BiDaaS Project, this Section develops several recommendations addressed to any manufacturing company willing to undertake Big Data projects and positions the I-BiDaaS solution within Big Data Value (BDV) reference model and Strategic Research and Innovation Agenda (SRIA).



## 5.1 Lessons learned, challenges and guidelines

I-BiDaaS project developed an integrated platform for processing and extracting actionable knowledge from Big Data in the manufacturing sector. Based on the challenges experienced and lessons learned through our involvement in I-BiDaaS, we propose a set of guidelines for the implementation of Big Data analytics in the manufacturing sector, with respect to the following concerns:

1. **Data storage and ingestion from various data sources and its preparation:** In a production line deploying with digital instruments, there are many devices which setup operating values, adjust and control parameters during the production processes. In many cases, data, which are used for the operation of the equipment, are overwritten because every day, large amount of data are generated, and often time to gather and analyse them is limited due to the fast rhythms of production. Furthermore, we need to consider fast internal changes due to production rescheduling, product quantities and component variations that can lead to variations in availability and data flow. Depending on whether we want to act on the quality of the production process or on the maintenance of the equipment, the first challenge is to understand how data will be ingested and managed from data sources over time and who will be able to access them. Furthermore, this aspect highlights the importance of breaking data silos by extracting the value of all data collected from several sources and levels and may be necessary to involve different departments belonging to the same or different organizations.
2. **Data cleaning:** A second important aspect is to understand which types of data can be useful for analysis. This implies the importance of data cleaning in order to identify incomplete, inaccurate and irrelevant parts of the generated dataset.
3. **Fabrication of realistic synthetic data for experimentation and testing:** Data are strictly confidential, so another challenge is to decide how data will be shared if external analysis is required. In this case, manufacturers need to evaluate the possibility of fabrication of realistic synthetic data for experimentation of the analytical models that will be developed and then to test the same models with anonymized real data. Furthermore, this aspect is also linked to the first. When many different types of information are gathered, the need may occur to analyse them in order to understand what is useful.
4. **Batch and stream analytics for increasing the speed of data analysis:** After collecting and analysing data, it is necessary to understand which Big Data technologies are most suitable for the specific identified business requirements. Batch and stream analytics cover all aspects, which may occur in real-world environments, including cases that require a deeper analysis of large amounts of data collected over a period of time (batch) or those that require velocity and agility for the events that we need to monitor in real or near real-time (streaming).
5. **Simple, intuitive and effective visualization of results and Interaction capabilities for the end-users:** The advanced visualisation tools which provide the insights, value, and operational knowledge extracted from available data,

need to consider both expert, and non-expert end-users. In the context of manufacturing activities, end-users can be grouped into three main categories: (i) manufacturers who have the relevant experience and current practices to innovate and improve, and offering the opportunity to validate and demonstrate the project, its approach and results across real contexts; (ii) intermediate users who are involved in data collection, data security, manual analysis, operational flows and required functionalities in order to innovate the production management processes; and (iii) operators who are usually employed at different levels in production processes, who need to timely visualise the data processing results.

## 5.2 Connection to BDV reference model and SRIA

The described solution for the defined manufacturing use cases can be contextualized within the BDV reference model defined in the BDV Strategic Research and Innovation Agenda (BDV SRIA). They contribute to the BDV reference model in the following ways. Specifically, regarding the BDV reference model horizontal concerns, we address:

- **Data visualization and user interaction:** by developing several advanced and interactive visualization solutions applicable in the manufacturing sector, as detailed in Sections 4.1.3 and 4.2.3.
- **Data analytics:** by developing data analytics solutions for the two industrial use cases in the manufacturing sector, as described in Sections 4.1.2 and 4.2.2. While the solutions may not correspond to state-of-the art advances in AI/machine learning algorithms development, they clearly contribute to revealing novel insights and best practices on how Big Data analytics can improve manufacturing operations.
- **Data processing architectures:** we develop architectures as shown in Figures 2 and 9 that are well-suited for manufacturing applications wherein both batch analytics (e.g., analyzing historical data) and streaming analytics (e.g., online processing of the data that correspond to a newly manufactured engine) are required.
- **Data protection and data management:** Real data were anonymized by CRF that manipulated and masked them after retrieving from an internal proprietary server.

Regarding the BDV reference model vertical concerns, we address the following:

- **Big Data Types and Semantics:** our work here is mostly concerned with structured sensory data, meta-data, and thermal images data (that corresponds to the Media, Image, Video and Audio data type according to the BDV nomenclature). The work also contributes to best practices in the generation of realistic synthetic data from the corresponding domain-defined meta-data, as well as a systematic way to assess the quality and usefulness of the generated synthetic data.

- **Communication and Connectivity:** the work describes innovative ways how to communicate with and retrieve data from an internal manufacturing company proprietary server, as described in Section 4 and outlined in Figure 1.

Therefore, in relation with BDV SRIA, the I-BiDaaS solution contribute to the following technical priorities: Data protection; Data Processing Architectures; Data Analytics; and Data Visualisation and User Interaction.

Furthermore, in relation with the BDVA SRIA priority areas in connection with Factories of the Future with EFFRA, we address the following dimensions:

- a) Excellence in manufacturing: Advanced manufacturing processes and services for zero-defect and innovative processes and products.
- b) Sustainable value networks: Manufacturing driving the circular economy
- c) Inter-operable digital manufacturing platforms: Supporting an ecosystem of manufacturing services.

In more detail, CRF use cases have been selected in order to develop innovative tools and solutions that may ensure better product quality towards zero-defect manufacturing. In particular, the existing production lines may be improved to maximise the quality of their product through the integration of solutions that exploit Big Data technologies. A better process efficiency can result in energy saving and cost reduction in the context of circular economy and allow to manufacturers to reach a high level of competitiveness and sustainability.

## 6 Conclusion

The increasing levels of digitalization in the manufacturing sector contribute to generate a large amount of data that often contain a high value of hidden information. This is due to the complexity of real processes that require several interconnected stages to obtain finished goods. Variables and parameters are set for the operation of each digital machine and just like we assembly components, we need to pull together data generated from different sources and levels, if we want to improve the quality of processes and products. I-BiDaaS developed an integrated platform, taking into consideration how complex data can be managed and how to help manufacturers who are not sufficiently enable to analyse complex datasets, by empowering them to easily utilize and interact with Big Data technologies.

## Acknowledgements

The research presented in this book chapter was undertaken in the framework of the project I-BiDaaS (“Industrial-Driven Big Data as a Self-Service Solution”) funded by the Horizon 2020 Programme under Grant Agreement 780787.

## References

- [1] B. Chen; J. Wan; L. Shu; P. Li; M. Mukherjee; B. Yin: Smart Factory of Industry 4.0: Key Technologies, Application Case, and Challenges, IEEE Access Volume 6, pages 6505 – 6519, (December 2017)
- [2] Klaus Schwab, The Fourth Industrial Revolution, 2016
- [3] Yadegaridehkordi, E., Hourmand, M., Nilashi, M., Shuib, L., Ahani, A., & Ibrahim, O. (2018). Influence of big data adoption on manufacturing companies' performance: an integrated DEMATEL-ANFIS approach. *Technological Forecasting and Social Change*, 137, 199-210
- [4] <http://www.ibidaas.eu/>
- [5] Zillner, S., Curry, E., Metzger, A., Auer, S., & Seidl, R. (Eds.). (2017). *European Big Data Value Strategic Research & Innovation Agenda*.
- [6] Arruda, D.: Requirements engineering in the context of big data applications. *SIG-SOFT Softw. Eng. Notes* 43(1), 1–6 (Mar 2018)
- [7] B. Nuseibeh and S. Easterbrook. Requirements engineering: A roadmap. In *Proceedings of the Conference on The Future of Software Engineering, ICSE '00*, pages 35–46, New York, NY, USA, 2000. ACM.
- [8] Horkoff, J., A.F.C.E.: Goal-oriented requirements engineering: an extended systematic mapping study. *Requirements Eng* 24(2019), 133-160 (2019)
- [9] Arruda, D.: Requirements engineering in the context of big data applications. *SIG-SOFT Softw. Eng. Notes* 43(1), 1-6 (Mar 2018)
- [10] NIST Big Data Public Working Group: Use Cases Requirements Subgroup: National Institute of Standards and Technology (NIST) Big Data Interoperability Framework: Volume 3, Use Cases and General Requirements. Tech. rep., Special Publication 1500-3 (2015)
- [11] I-BiDaaS Consortium: D1.3: Positioning of I-BiDaaS (2018), available at: <http://www.ibidaas.eu/sites/default/files/docs/Ibidaas-d1.3.pdf>
- [12] Arapakis, Ioannis, et al. "Towards specification of a software architecture for cross-sectoral big data applications." *2019 IEEE World Congress on Services (SERVICES)*. Vol. 2642. IEEE, 2019.
- [13] Paszke et al., "PyTorch: An Imperative Style, High-Performance Deep Learning Library", 2019, available at: <https://arxiv.org/abs/1912.01703>
- [14] <https://www.docker.com/>
- [15] J. Gua et al.: Recent Advances in Convolutional Neural Networks, *Pattern Recognition*, Volume 77, May 2018, Pages 354-377
- [16] G. Huang, Z. Liu, L. van der Maaten, K. Q. Weinberger: Densely Connected Convolutional Networks, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700-4708
- [17] K. He, X. Zhang, S. Ren, J. Sun: Deep Residual Learning for Image Recognition, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778
- [18] Visualization with TensorBoard: <https://www.tensorflow.org/tensorboard>
- [19] J. Snoke, G. Raab, B. Nowok, C. Dibben, A. Slavkovic: General and specific utility measures for synthetic data, arXiv:1604.06651 (2017), available at <https://arxiv.org/pdf/1604.06651.pdf>
- [20] M. Gupta, J. Gao, C. C. Aggarwal and J. Han, "Outlier Detection for Temporal Data: A Survey," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 9, pp. 2250-2267, Sept. 2014, doi: 10.1109/TKDE.2013.184.
- [21] F. Petitjean; G. Forestier; G. I. Webb; A. E. Nicholson; Y. Chen; E. Keogh: Dynamic Time Warping Averaging of Time Series Allows Faster and More Accurate Classification, 2014 IEEE International Conference on Data Mining