6

Smart Data and the Industrial Internet of Things

Christian Beecks, Hassan Rasheed, Alexander Grass, Shreekantha Devasya, Marc Jentsch, José Ángel Carvajal Soto, Farshid Tavakolizadeh, Anja Linnemann and Markus Eisenhauer

Fraunhofer Institute for Applied Information Technology FIT,
Schloss Birlinghoven, Sankt Augustin, Germany
E-mail: Christian.Beecks@fit.fraunhofer.de

Abstract

Many modern production processes are nowadays equipped with cyber-physical systems in order to capture, manage, and process large amounts of sensor data including information about machines, processes, and products. The proliferation of cyber-physical systems (CPS) and the advancement of Internet of Things (IoT) technologies have led to an explosive digitization of the industrial sector. Driven by the high-tech strategy of the federal government in Germany, many manufacturers across all industry segments are accelerating the adoption of cyber-physical system and IoT technologies to gain actionable insight into their industrial production processes and finally improve their processes by means of data-driven methodology. In this work, we aim to give insights into our recent research regarding the domains of Smart Data and Industrial Internet of Things (IIoT). To this end, we are focusing on the EU projects MONSOON and COMPOSITION as examples for the Public-Private Partnership (PPP) initiatives Factories of the Future (FoF) and Sustainable Process Industry (SPIRE) and show how to approach data analytics via scalable and agile analytic platforms. Along these analytic platforms, we provide an overview of our recent Smart Data activities and exemplify data-driven analysis of industrial production processes from the process and manufacturing industries.

6.1 Introduction

Many modern production processes are nowadays equipped with cyber-physical systems in order to capture, manage, and process large amounts of sensor data. These sensor data include information about machines, processes, and products and are encountered in form of data streams. These data streams from the production site are then frequently integrated into cloud-based solutions by means of Internet of Things technologies in order to allow comprehensive data-driven investigations and process optimizations.

The proliferation of cyber-physical systems and the advancement of IoT technologies have led to an explosive digitization of the industrial sector. Driven by the high-tech strategy of the federal government in Germany, many manufacturers across all industry segments are accelerating the adoption of cyber-physical systems and Internet of Things technologies in order to gain actionable insight into industrial production processes and finally improve these processes by means of data-driven methodology.

The IoT is one of the key enabler for intelligent manufacturing and production. It facilitates the intelligent connectivity of smart embedded devices in factories and shop floors. Endowing the manufacturing and production site with technologies from the IoT, which is then also referred to as the IIoT, has become a technical prerequisite for a sustainable and competitive industrial production of the future.

Digitizing the industrial sector with cyber-physical systems, Internet of Things technologies, cloud computing services, and Smart Data analytics leads to the fourth industrial revolution, which is denoted as Industry 4.0. The importance of strengthen the European industry to become more sustainable and competitive is also taken into account by the European Commission. Within the EU Framework Programme for Research and Innovation the two Public-Private Partnership (PPP) initiatives Factories of the Future (FoF) and Sustainable Process Industry (SPIRE) aim to (i) help EU manufacturing enterprises to adapt to global competitive pressures by developing the necessary key enabling technologies across a broad range of sectors and (ii) support EU process industry in the development of novel technologies for improved resource and energy efficiency.

Turning industrial Big Data into structured and useable knowledge is one of the major data-centric challenges for enhancing production processes. Integrating data from heterogeneous systems and gaining insight into voluminous amounts of streaming sensor data with high variety and velocity requires scalable methods and techniques. Structuring knowledge in a way that it can be used to manage and improve industrial production processes is one of the objectives of Smart Data analytics. By improving Big Data to a higher degree of quality, Smart Data analytics aims to understand the following aspects:

- Purpose: What problem to solve with the data?
- People: Who is involved?
- Processes: What are the surrounding processes?
- Platform: Which IT infrastructure is necessary for realization?

The aforementioned aspects are also referred to as the 4Ps of Smart Data. They indicate the information to be gathered in addition to the sensor data from the production site in order to get a more complete understanding about the data and its surrounding entities. It is obvious that addressing the 4Ps within the Smart Data analytics process strongly relies on user-centered methods since many of the required information need to be discovered from non-documented data.

The Fraunhofer Institute for Applied Information Technology FIT has been conducting research and development on user-friendly smart solutions that blend seamlessly in business processes for about 30 years and has a strong experience in digitization, Industry 4.0 projects and IoT solutions. Having about 160 researchers with different scientific background, the Fraunhofer Institute for Applied Information Technology FIT is organized into five research departments:

- The User-Centered Computing department develops IT systems and technologies that focus on their users throughout their complete life cycle. Current work focuses on usability engineering, web compliance, and accessibility.
- The Cooperation Systems department develops and evaluates groupware and community systems for virtual teams and organizations. Our work on hardware and software of Mixed and Augmented Reality systems focuses on support for cooperative planning tasks.
- The Life Science Informatics department designs and implements complex biomedical information systems and creates novel software solutions for manufacturers and users in health care, biotechnology, drug research and social services. Focal areas are image-based navigation systems, information-intensive optical instruments, visual information analysis, multi-parametric molecular sensor technology and diagnostics as well as bio-analogue analysis of changing images.
- The Risk Management and Decision Support department offers decision and process support for application domains whose processes can be

characterized by their high level of complexity as well as their weak determination of process structures.

- The Fraunhofer Project Group Business & Information Systems Engineering, located in Augsburg and Bayreuth, has proven expertise at the interface of Financial Management, Information Management and Business & Information Systems Engineering. The ability to combine methodological know-how at the highest scientific level with a customer-focused and solution-oriented way of working, is our distinctive feature.

As part of User-Centered Computing department, the User-Centered Ubiquitous Computing group develops systems providing effective personal assistance that dynamically respond to user demands and at the same time adapt to new work practices. The group is focusing on the application domains Industry 4.0, Smart Cities and Energy Efficiency/Smart Grids and approach novel applied solutions via methods from the domains User-Centered Design, Internet of Things Platforms, and Smart Data.

In this chapter, we aim to give insights into our recent research into the domains of Smart Data and Industrial Internet of Things. To this end, we are focusing on the following EU projects:

- The MONSOON (MOdel based coNtrol framework for Site-wide OptimizatiON of data-intensive processes) project aims to establish a data-driven methodology to support the identification and exploitation potentials by applying multi-scale model based predictive controls in production processes. It offers an integrated real-time and dependable infrastructure easing in improving the efficient use and re-use of raw resources and energy across plant- and site-wide applications in heterogeneous and distributed production environments. EU funds it under SPIRE (Sustainable Process Industry through Resource and Energy Efficiency) research project that aims to develop an infrastructure in support of the process industries.
- The COMPOSITION (Ecosystem for COllaborative Manufacturing PrOceSses – Intra- and Interfactory Integration and AutomaTION) project has two main goals: The first goal is to integrate data along the value chain inside a factory into one integrated information management system (IIMS) combining physical world, simulation, planning and forecasting data to enhance re-configurability, scalability and optimisation of resources and processes inside the factory. The second goal is to create a (semi-)automatic ecosystem, which extends the local IIMS concept to

a holistic and collaborative system incorporating and inter-linking both the Supply and the Value Chains. The COMPOSITION project is funded under the Factories of the Future PPP.

In conjunction with both EU projects mentioned above, EXCELL is a twinning project addressing Big Data applications for cyber-physical systems in production and logistics Networks. The consortium of academics from Hungary, Great Britain, Belgium and Germany expands the scientific activities through central publications and active participation in scientific discourses. Priority Research Fields (PRFs) define the topic areas in which the partners work closely together to mutually train, support and empower each other with their knowledge and expertise. PRFs are for example cyber-physical systems and human system interaction, business-based Internet of Things and services, as well as data mining and data interoperability.

In the remainder of this chapter, we will first describe our research activities and results with respect to the EU project MONSOON, which is an example for the process industry, in Section 6.2. Afterwards, we will continue with the EU project COMPOSITION, which is an example for the manufacturing/discrete industry, in Section 6.3. We finally conclude this chapter in Section 6.4.

6.2 Process Industry

6.2.1 Introduction

The process industry is characterized by intense use of raw resources and energy, and thus represents a significant share of European industry in terms of energy, resources consumption and environment impact. In this area, even a small optimization can lead to high absolute savings, both economic and environmental. Predictive modelling techniques can be especially effective in optimization of production processes. However, the application of these techniques is not straightforward. Predictive models are built using the data obtained from production processes. In many cases, process industries must invest in the monitoring and data integration as well as in the development and maintenance of the underlying infrastructure for data analytics. Many other obstacles are also present, e.g., interoperability issues between software systems in production, difficulties in the physical monitoring of the production parameters, problems with the real-time handling of the data, or difficulties in defining relevant Key-Performance Indicators (KPIs) to support management. Therefore, the deployment of such predictive functions in production with reasonable costs requires consolidation of the available resources into shared cloud-based technologies. In the case of more flexible production environments, approaches that are even more significant are possible, such as the reinvention or redesign of the production processes. However, this is not applicable to major, capital-intensive process industries. In this case, the integration of innovations in the established production processes can be fundamental in their transformation from resource-consuming production into the “circular” model.

6.2.2 Reference Architecture

The high-level conceptual view of the reference architecture that is developed within the scope of the project MONSOON is depicted in Figure 6.1.

The platform is able to inter operate with the heterogeneous existing systems deployed in process industries at different layers of the SCADA pyramid (Control, Supervision, Management, Enterprise). It includes sensors or controllers (PLCs), SCADA (Supervision control and data acquisition), Management Information Systems (MES) and Enterprise Resource Planning (ERP). There are two main components of the architecture. The Real-time Plant Operations Platform deployed on-site and supports data collection, storage and interaction with the production systems respecting relevant constraints and satisfying data-intensive conditions. The Cross-Sectorial Data

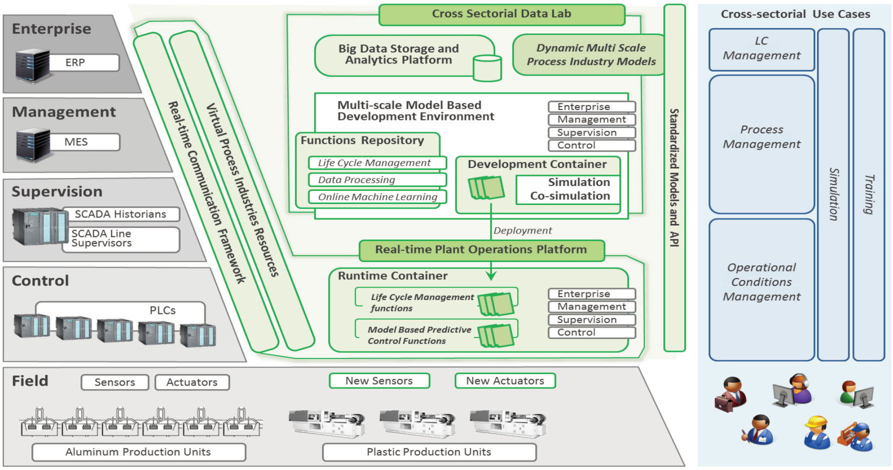

Figure 6.1 MONSOON Reference Architecture.

Lab supports the development of new dynamic model base multi-scale controls. All the relevant data from the production site are transferred to the Data Lab where it is stored and processed for optimization of production process. To validate and demonstrate the results, two real environments are used within the project: an aluminium plant in France and a plastic factory in Portugal. We have identified two main use cases for both domains.

For the aluminium sector, we focused on production of the anodes (positive electrodes) used in aluminium extraction by electrolysis. The first use case was targeted to predictive maintenance, where the main objective was to anticipate the breakdowns and/or highlight equipment/process deviations that affect the green anode final quality (e.g., anode density). The second use case dealt with the predictive anode quality control, where the goal was to identify bad anodes with a high level of confidence and scrap them to avoid sending them to the electrolysis area.

For the plastic domain, the use cases are from the area of production of coffee capsules, produced in large quantities. In this type of production, it is important to produce the correct diameter and height of the coffee capsules and to make sure that the holes at the bottom of the capsules are formed properly. Moreover it is also expected to predict the failures of molding machines and their stoppages based on the process parameters and sensor measurements during molding processes. While the data analysis process for the plastic domain is described in Section 6.2.6, we provide a short description of main components of the MONSOON platform along with their interfaces in the next sections.

6.2.3 Plant Operational Platform

The functional view of the architecture of the Plant Operations Platform is presented in Figure 6.2. It acts as an advanced semantic factory service bus and is in-charge of interacting with existing production systems deployed in a plant. The Plant Platform IT infrastructure and its associated Real-time Data Integration layer collect the operational raw data from the plant's systems necessary to the execution of the predictive functions. The acquired operational raw data and associated relevant information is also routed to the data lab where it is stored and used for analytics.

6.2.3.1 Real-time communication framework

It configures the dependable real-time communication infrastructure necessary to support operations of prediction functions. The *Monitoring Tools*

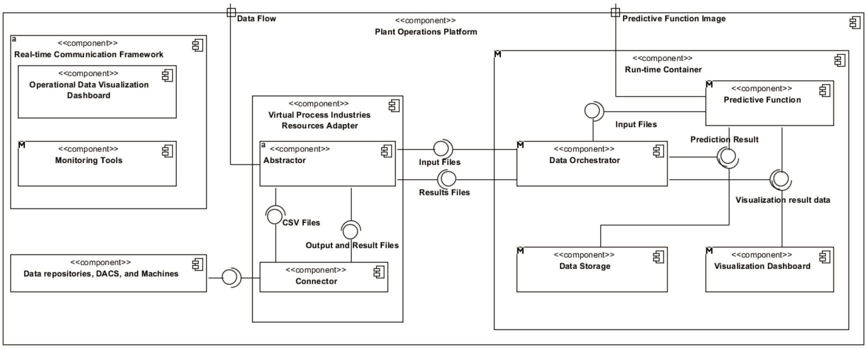

Figure 6.2 Functional View of Plant Operational Platform.

exploit and integrate existing solutions for real-time networking and QoS management and perform continuous (passive/active) monitoring of plant-wide process industry resources ensuring that communication-related malfunctions are properly detected. The *Operation Data Visualization Dashboard* provides a web user interface where operational managers can configure various real-time visualizations of operational data and monitor the deployed predictive functions. The visualized data can include operational data from the plant environment or predictions from the predictive functions executed in the Run-time Container.

6.2.3.2 Virtual process industries resources adapter

The main function of the Virtual Process Industries Resources Adapter (VPIRA) is data integration, mediation and routing. The *Connector* allows the integration of data from various SCADA, MES and ERP systems deployed on the plant site. It ensures that all heterogeneous process industry resources and systems are easily accessed and managed. The *Abstractor* is a distributed and scalable data flow engine aiming for routing integrated data to multiple destinations, e.g., run-time container or data lab. Routing of data from the source to target connectors can be dynamic depending on the type of data or actual content. The data flows can be re-configured in a flexible way, connecting multiple sources to the multiple targets, overcoming any data heterogeneity problems. Besides the flexible configuration interface, the Virtual Process Industries Resources Adapter provide a flexible programming interface to simply implement connectors or processors for new types of data sources and formats.

6.2.3.3 Run-time container

The Run-time Container executes the model based predictive functions and life-cycle management functions within the overall plant infrastructure. It ensures proper deployment and execution of predictive functions developed by means of the data lab, hence it manages all aspects of predictive functions life cycle. It is composed into four sub-components as described below:

- *Data Orchestrator*: coordinates the data flow between different components, such as transmit input data, store prediction result, and pass visualization result data to relevant components.
- *Predictive Function*: exports predictive function image from Function Repository and instantiate the execution of predictive function that perform real-time scoring of input operational data. It performs all operations required for the pre-processing of raw data into inputs for the specific predictive function and into process prediction output.
- *Data Storage*: stores the prediction results into a scalable database. The prediction results are also sent to the Operational Platform systems and the data lab for combining these real-time results with historical data analysis.
- *Visualization Dashboard*: displays prediction results and generates feedback instructions or alerts towards plant's systems to inform/warn the site operators to adjust the process regulation parameters.

6.2.4 Cross Sectorial Data Lab Platform

The Data Lab provides a collaborative environment where high amounts of data from multiple sites, and possibly from multitude of industry sectors, are collected, stored and processed in a scalable way. It enables multidisciplinary collaboration of experts allowing teams to jointly model, develop and evaluate distributed controls in rapid and cost-effective way. The Data Lab eases the definition of predictive control and life cycle management functions, allowing to work in a simulated environment or to exploit co-simulation by mixing stored data with data flowing in real-time from the real systems.

The Data Lab thus supports data science and automation experts interested to optimization and scheduling aspects by providing the suitable environment to mine, process, re-play production data. It allows modelling of the whole production process across the SCADA layers including the specification of the data dictionary of all inputs and outputs of the processing steps and their relations to the overall KPIs. The semantic models capture the site knowledge base for given application cases and used data analytics

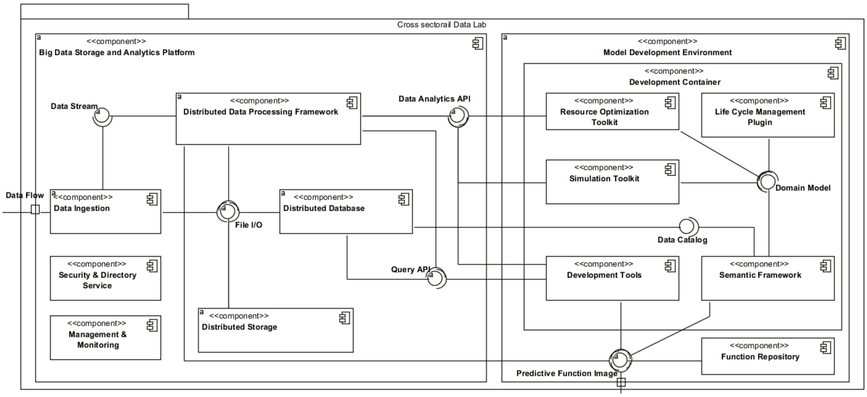

Figure 6.3 Functional View of Cross Sectorial Data Lab Platform.

methods allowing generalization of cases to existing good practices and transfer of the knowledge by adaptation of cases to new environment/site. The main outcome of the Data Lab is typically a single or multiple new predictive functions and life cycle management controls ready to be deployed in the Runtime Container of the Plant Operations Platform.

The components of the Cross Sectorial Data Lab are shown in Figure 6.3 and explained in the sections below.

6.2.4.1 Big data storage & analytics platform

The Big Data Storage and Analytics Platform provides resources and functionalities for storage as well as batch and real-time processing of the operational data from multiple site characterized as Big Data. The platform combines and orchestrates existing technologies from the Big Data and Analytic landscape and sets a distributed and scalable run-time infrastructure for the developed data analytics methods. It provides main integration interfaces between the site Operational Platform and the cloud Data Lab platform and the programming interfaces for the implementation of the data intensive analytics methods. The Big Data Storage and Analytics Platform consist of the following sub-components:

- *Distributed Storage*: provides a reliable, scalable file system with similar interfaces and semantics to access data as local file systems.
- *Distributed Database*: provides a structured view of the data stored in the platform using the standard SQL language, and supports standard RDBMS programming interfaces such as JDBC for Java or ODBC for Net platforms.

- *Distributed Data Processing Framework*: allows the execution of applications in multiple nodes in order to retrieve, classify or transform the arriving data. The framework provides Data Analytics APIs for processing large datasets via parallel and distributed computations.
- *Data Ingestion*: implements an interface for real-time communication between the Data Lab and Operation platforms. It also supports batch uploading of the historical data between the Data Lab and Operation platform.
- *Security & Directory Service*: provides user management and content authorization capabilities for the platform services.
- *Management & Monitoring*: provides the management, monitoring and provisioning of the platform services on the hosted environment.

6.2.4.2 Model development environment

The Model Development Environment provides tools and interfaces that cover the whole life cycle of planning, implementation, testing, validation and deployment of predictive functions and life-cycle management controls into the plant production supporting simulation/co-simulation features.

- *Development Tools*: provide the main collaborative and interactive interface for data engineers, data analysts and data scientists to execute and interact with the data processing workflows running on the Data Lab platform. Using the provided interface, data scientists can organize, execute and share data, and code and visualize results without referring to the internal details of the underlying Data Lab run-time infrastructure. The interface is integrated in form of analytical "notebooks" where different parts of the analysis are logically grouped and presented in one document. These notebooks consist of code editors for data processing scripts and SQL queries, and interactive tabular or graphical presentations of the processed data.
- *Semantic Modelling Framework*: provides a common communication language between domain experts, stakeholders and data scientists. A collaborative web interface is provided for the creation and sharing of semantic models in order to use the knowledge expressed in such models for the optimization of the production processes in the Simulation and Resource Optimization Framework.
- *Simulation Toolkit*: supports validation and deployment of predictive functions in order to optimize overall KPIs defined for the production process. The estimation of overall impacts can be used to test various "what if" scenarios, or for the automatic discrete optimization of the

production process by finding the optimal combination of predictive functions for various process phases.

- *Resource Optimization Toolkit:* optimizes the production process based on various indicators representing the performance of manufacturing process of the plant leveraging process data and knowledge extracted from analytics methods.
- *Life-Cycle Management Plugin*: serves as multi-disciplinary, transversal tool to evaluate environmental performance of a given production process for life-cycle environmental indicators, such as Global Warming Potential and Total Energy Requirement.

6.2.4.3 Function repository

The Function Repository provides a storage for predictive functions together with all settings required for the deployment of predictive functions, where they are available for production deployment or for the simulations and overall optimization of the production processes. The predictive functions are packaged as container images so that entire predictive function pipeline (including pre-processing and task specific evaluation) can be implemented within a virtualized container.

6.2.5 Deployment

The Data Lab Platform promises to combine and orchestrate existing technologies and open source frameworks from the Big Data landscape to establish a distributed and scalable run-time infrastructure for the data analytics methods. We present in Figure 6.4 the mapping of the platform components

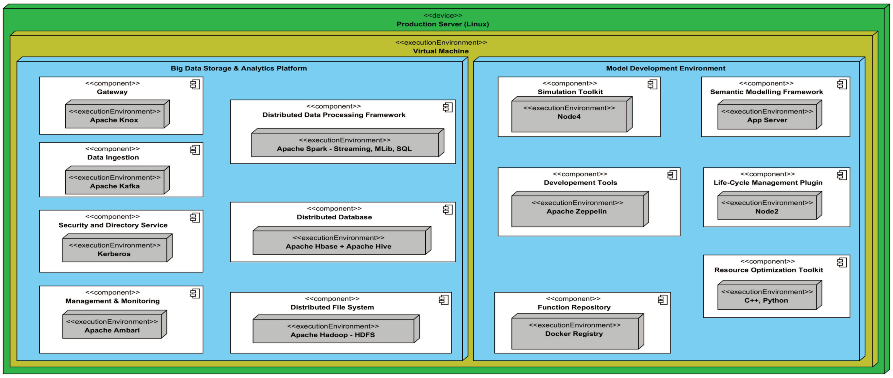

Figure 6.4 Components Mapping to Open-source Technologies.

to existing and emerging open source technologies selected and used during the initial deployment.

The initial deployment was performed with multiple virtual machines on an in-house physical infrastructure. It turned out that the overall deployment time and configuration management is the most critical aspect in realizing and operationalizing such a platform. It would be optimal to devise a uniform deployment strategy taking into account different deployment options for the platform such as on-premises, cloud/external provider or hybrid. It has also been learned that different demonstrative and use-case scenarios in both aluminium and plastic domains pose different infrastructure and data requirements. Hence, it is useful to define different deployment pipelines or modes for the platform where the right set of platform services are deployed and orchestrated accordingly instead of full stack deployment. Towards this goal, the Big Data Storage and Analytics Platform has been containerized to adapt a common deployment ground with the objective of easing the usage of common platform technologies and make integration with other services or applications easy. The containerization based on Docker framework is depicted in Figure 6.5.

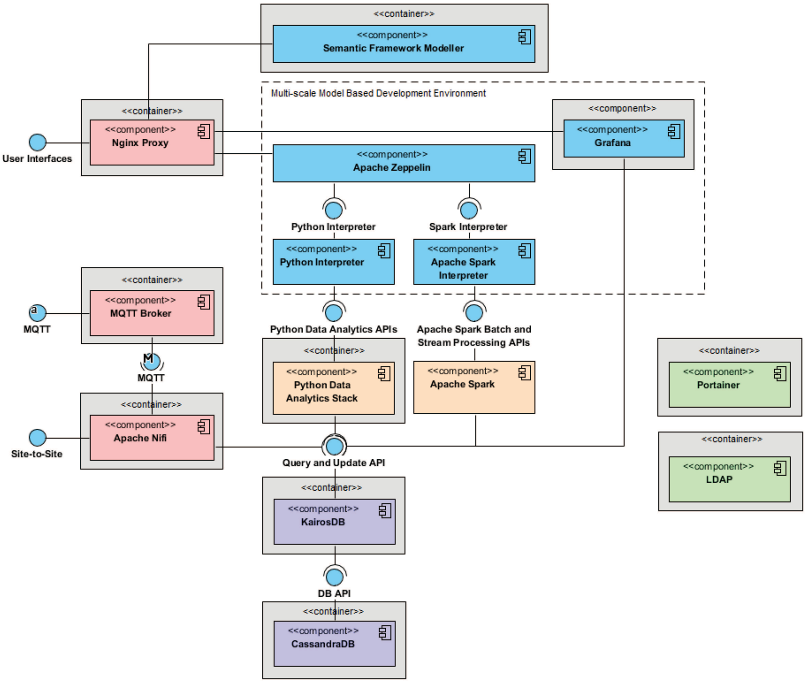

Figure 6.5 Containerization of Big Data Storage and Analytics Platform.

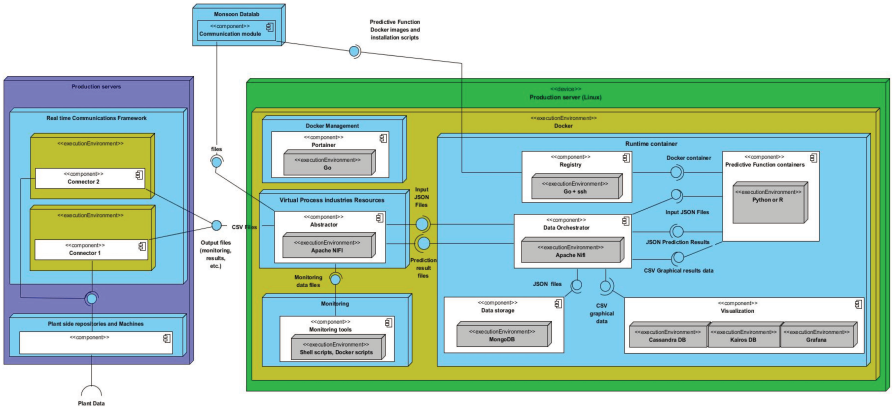

Figure 6.6 Deployment view of Plant Operational Platform.

The deployment of the Site Operational Platform with open source technologies mainly for Virtual Process Industries Resources Adaptor and Run-time Container is finally illustrated in Figure 6.6. It shows how predictive functions can be applied in factorial settings.

6.2.6 Data Analysis

Data analysis in process industries mainly aims to reduce the wastage of time, resource and energy during production processes. This can be achieved by several means: avoiding equipment stoppages, maintaining optimum configurations, early detection of a chain of events causing an anomaly etc. Data analysis is simplified by the components of the *Cross Sectorial Data Lab* which provides a single platform for data fetching, accessing and artefact development. The data collected from the plastic molding machines are stored in the *Big Data Storage & Analytic Platform.* These data are used by the data scientists and the process experts for exploratory analysis in order to gain initial insights. The collaborative interface provided by the platform is used simultaneously by the process expert and the data scientists. The findings from the exploratory analysis is used as the basis for modelling the process leading to the development of predictive functions. These functions are stored in the *Function Repository* which are deployed later in factory premises for real-time predictions. Although this process is generic enough to be applied in any kind of industrial environment, we shall limit our discussion to the plastic industry.

The objective of data analysis in the plastic industry is to anticipate the breakdowns and/or highlight equipment/process deviation that impacts the injection molding process and therefore to improve the quality of the produced coffee capsules. In general, there are two areas where waste parts can occur in plastic injection molding process: the molding tool and the molding process. During the long-term production of the coffee capsules, parameters of the injection molding process can slightly change due to various changes of the environment (temperature and humidity in the factory, deviations in the energy supply system, heating of oil temperature, deviations in the quality of the plastic granules, wearing of machine parts). The aim is to monitor technical parameters of the molding machine and raise an alarm if the deviation is increasing over the defined values. These long-term changes can also cause the stoppage of molding machines. Which in turn causes reduction of produced capsules. In addition, few of the initial cycles after restart are wasted during the calibration process producing defective capsules.

6.2.6.1 Data description

Two kinds of data have been collected in the first year from GLN site during the production of coffee capsules. The first data set is collected automatically from a Euromap63 interface recorded on molding machines and the second data set is collected during the experiments conducted by a process expert during their visit to the production site.

The first data set is unlabeled and contains sensor measurements of several coffee capsule production cycles. Each cycle lasts almost 7 seconds, except if it causes a breakdown. The data set has a total of 88 attributes representing temperatures, time taken for different stages, pressure, cylinder positions etc. All data were directly monitored by the injection molding machine and stored there. Of them, only 12 (heating belt temperatures, maximum cycle pressure, coolant temperatures, residual melt cushion, plastification time) are proposed as useful, and, particularly, their ranges/deviations over intervals instead of their values themselves are suggested to serve as explanatory variables.

The second data set [1] is manually labelled and comprises information about 250 production cycles of coffee capsules from the injection molding machine and their quality information. It contains 36 attributes reflecting the machine's internal sensor measurements for each cycle. These measurements include values about the internal states, e.g. temperature and pressure values, as well as timings about the different phases within each cycle. In addition, we also take into account quality information for each cycle, i.e., the number of non-defect coffee capsules which changes throughout individual production cycles. The quality of each capsule is inspected by the domain expert in different aspects. i.e. the capsules have permissible range of height and base diameter. Also each capsule should have uniform thickness and should not have holes. If any of these expectations are not met, the capsule is considered to be defective. If the number of produced high quality coffee capsules is larger than a predefined threshold, we label the corresponding cycle with high.quality, and otherwise we assign the label low.quality. The decision about the quality labels was made by domain experts.

Exploratory analysis is performed on the unlabeled data in order to discover hidden insights. On the other hand, basic machine learning algorithms are applied to the labelled data to classify the cycles based on their quality. In the upcoming subsections we discuss these two different approaches on these data sets.

6.2.6.2 Preliminary trend analysis of unlabeled data

The main aim of the preliminary analysis of is to get some initial overall insights that might be interesting for the process experts to be further analyzed. The first step was to understand the attributes and their correlations. This was followed by visual exploration of data with manual inspection followed by clustering the data to find significant relation between different cycles. Considering the huge amount of data generated by sensors, clustering usually takes lots of time. One strategy is to use the computation powers of the Data Lab clusters to perform these operations faster. If the algorithms for exploratory data analysis are deployed in the Data Lab, domain experts and data scientists can use the results simultaneously to get actionable insights.

One of the insight was repeating set of parameters in the data. This was found by using matrix profiles. A pattern obtained by applying matrix profiles is the decrease in plastification time and at the same time, increase in cycle time. Plastic domain expert cross checked these patterns and found out that this happens whenever there is an equipment stoppage due to lack of lubrication. Though the characteristics of these incidents are known, early prediction of the possible stoppage has not been found out with data analysis. The corresponding patterns are shown in Figure 6.7.

Preliminary trend analysis helps us to extract the knowledge hidden in voluminous unlabelled data sets. This process can be automated to get the best results in minimum time. In addition, in the MONSOON project we include many stakeholders such as process experts, machine supervisors and ground workers to actively contribute to the production process optimization.

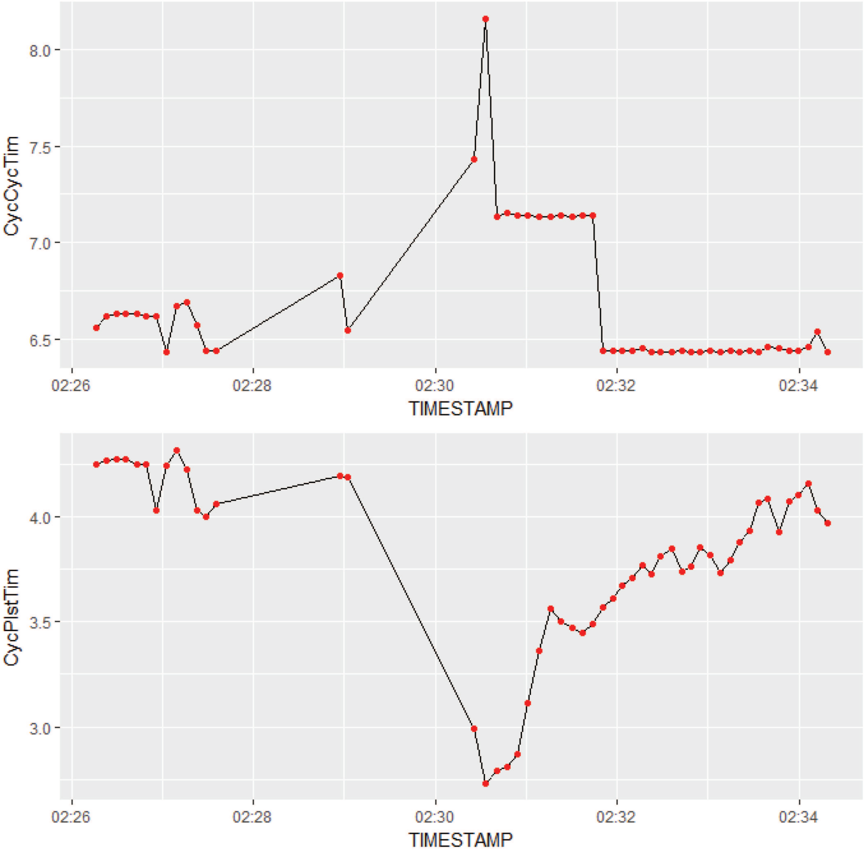

Figure 6.7 Increase in cycle time and decrease in plastification at the same time. The same pattern has repeated multiple times in the unlabelled set of plastic data. CycCycTim is cycle time and CycPlstTim is plastification time.

This is achieved with the help of a centralized Big Data analytics platform. On deploying the knowledge discovery algorithms in the Big Data analytics platform, the stakeholders can give live feedbacks. The data scientists further use these feedbacks for deriving conclusions. This is an ongoing work as part of the project.

6.2.6.3 Machine learning for labelled data

The goal of the machine learning process is to classify the injection molding cycles to high and low-quality cycles. As discussed earlier, the cycles are labelled as *high.quality* or *low.quality* based on the number of defective capsules produced in a cycle beyond a threshold defined by the process expert.

The initial dataset is pre-processed as follows. The labelled data is first centred and scaled. Later, the number of attributes is reduced by excluding the ones with near zero variance. Principal Component Analysis is applied to the remaining attributes to get the projection of data in reduced number of dimensions.

Basic classification algorithms, namely, k-Nearest Neighbour, Naïve Bayes, Classification and Regression Trees (CART), Random Forests and Support vector Machines (SVM) are investigated on the pre-processed data. SVM is investigated both with linear and RBF kernels. The performance of the models are measured in terms of balanced accuracy, precision, recall and F1 scores. K-fold cross validation is used to evaluate the performance. The number of folds is set to 5 and the number of repetitions is set to 100. We used 80% of the dataset is for training and 20% for testing. This investigation is performed via the CARET package in the programming language R. The results of our performance evaluation are summarized in Table 6.1.

From the table above, we see that all predictive models reach an accuracy of minimum 63%. The highest accuracy is achieved by the k-Nearest Neighbour classifier predicting the correct quality labels for more than 69% of the data.

Albeit these results were satisfying, these algorithms cannot be deployed straight away as the data used for this performance evaluation has been manually labelled by the experts. In the situations where the capsules are produced in millions per day, it is wiser to use the automatically labelled data for training the models and deploy them afterwards. One approach is to use the decision of the visual inspection systems in order to label the data. But this is not trivial since there is no one to one mapping between the optical inspection systems data and the actual cycle data. This is because multiple capsules belonging to different cycles and machines are passed to

Table 6.1 Classification results of different predictive models

	Balanced Accuracy	Precision	Recall	F1 score
k-NN	0.697	0.638	0.686	0.657
Naïve Bayes	0.643	0.604	0.563	0.578
CART	0.637	0.595	0.566	0.573
Random Forest	0.653	0.619	0.570	0.589
SVM (linear)	0.632	0.626	0.488	0.540
SVM (RBF)	0.663	0.643	0.563	0.594

the automatic visual inspection system at once making it harder to identify individual cycles belonging to a particular machine.

6.2.7 Summary

In this section, we have presented our recent research activities within the scope of the EU project MONSOON: As an example for the process industry, we have described the overall reference architecture facilitating cross-sectorial data analytics. As part of our ongoing work, we have also highlighted the analysis of sensor data arising from the plastic industry sector. In the following section, we will focus on the manufacturing industry.

6.3 Manufacturing/Discrete Industry

6.3.1 Introduction

As an example for the manufacturing industry, we focus on the EU project COMPOSITION. This project addresses the requirements of modern production processes, which stress the need of greater agility and flexibility leading to faster production cycles, increased productivity, less waste and more sustainable production. At the factory level, decisions need to be supported by detailed knowledge about the production process and its interplay with external entities. Unfortunately, historical and live data that generates this knowledge is becoming more and more distributed and few solutions are available that can easily tackle the implied challenges. Moreover, factories are becoming less isolated in the productive tissue of nations and several suppliers and third-party service providers need to be contacted and coordinated to implement decisions taken at the factory level.

In such a worldwide and dynamic environment, the ability of automatizing the preliminary coordination and negotiation activities involved in setting up supply chains for specific needs, in an open marketplace-like fashion, could greatly improve the ability of factories to quickly react to external challenges and driving forces.

6.3.2 Intra-factory Interoperability Layer Part of the COMPOSITION Architecture

In this chapter, we will address the COMPOSITION architecture in the data analytics context. The intra-factory interoperability layer has two main goals: the first one is to provide an infrastructure to combine distributed

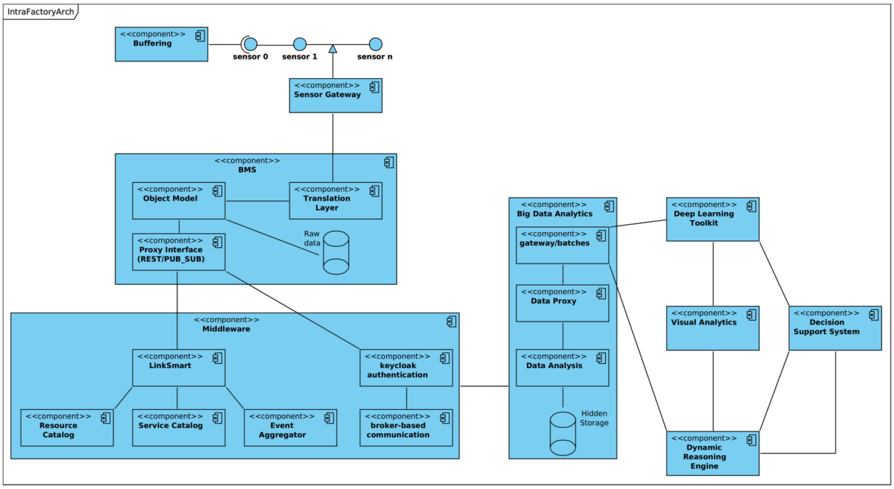
IntraFactoryArch
<<component>> Buffering
sensor 0
sensor 1
sensor n
<<component>> Sensor Gateway
<<component>> BMS
<<component>> Object Model
<<component>> Translation Layer
<<component>> Proxy Interface (REST/PUB_SUB)
Raw data
<<component>> Middleware
<<component>> LinkSmart
<<component>> keycloak authentication
<<component>> Resource Catalog
<<component>> Service Catalog
<<component>> Event Aggregator
<<component>> broker-based communication
<<component>> Big Data Analytics
<<component>> gateway/batches
<<component>> Data Proxy
<<component>> Data Analysis
Hidden Storage
<<component>> Deep Learning Toolkit
<<component>> Visual Analytics
<<component>> Decision Support System
<<component>> Dynamic Reasoning Engine

Figure 6.8 Intra-factory interoperability layer components and dependencies.

data in the integrated information management system and to do data analytics, the second one is to ensure the conformity between communications among interconnected components. Figure 6.8 shows the relevant part of the architecture.

The components of the architecture are introduced and described in the following:

- The *BMS* is provided by a project development stakeholder and is the translation layer providing shop floor connectivity from sensors to the COMPOSITION system. Raw data storage is added for offline debug purposes.
- The *Middleware* is the main recipient in which the interoperability of single components act.
- *LinkSmart* is a well-known middleware solution per se and is customized to satisfy the requirements of the COMPOSITION project. LinkSmart comprises the following components:
    - The *Service Catalog* works as service index and provides security information for service intercommunication.
    - The *Event Aggregator* parses messages to ensure homogeneity in data streams.
    - *Keycloak* is a virtual layer that ensures authorization and authentication. Like all security related measures, it is deployed by the Security Framework.

  - ○ The *broker-based intra-factory communication system* manages all internal communication.

- The *Big Data Analytics* component provides Complex Event Processing (CEP) capabilities for the data provided by the intra-factory integration layer
- The *Hidden Storage* is an optional storage not accessible from the outside in which aggregated data are stored for debug purposes, i.e. re-bootstrapping already trained artificial neural networks belonging to the Deep Learning Toolkit and to the Dynamic Reasoning Engine.
- The *Visual Analytics* component is the reporting interface of the Decision Support System and Simulation and Forecasting Toolkit.
- The *Dynamic Reasoning Engine* is part of the Simulation and Forecasting Toolkit.
- The *Decision Support System* uses process models to guide the production process.

Having a fist overview of the components of the COMPOSITION project and their dependencies, we continue with describing our approach to smart data analysis in the following section.

6.3.3 The Complex-Event Machine Learning methodology

Manufacturing in assembly lines consist of a set of hundreds, thousands or millions of small discrete steps aligned in a production process. Automatized production processes or production lines thereby produce for each of those steps small bits of data in form of events. Although the events possess valuable information, this information loses its value over time. Additionally, the data in the events usually are meaningless if they are not contextualized, either by other events, sensor data or process context. To extract most value of the data, it must be processed as it is produced, to be more precise in real-time and on demand. Therefore, in case of Big Data Analyses we propose the usage of Complex-Event Processing for the data management coming from the production facilities. In this manner, the data is processed in the moment when it is produced, extracting the maximum value, reducing latency, providing reactivity, giving it context and avoiding the need of archiving unnecessary data.

The Complex-Event Processing service is provided by the LinkSmart® Learning Agent (LA). The LA is a Stream Mining service that provides the utilities to manage real-time data for several purposes. On the one hand, the LA provides a set of tools to collect, annotate, filter, aggregate, or cache the real-time data incoming from the production facilities. This set of tools facilitates the possibility to build applications on top of real-time data. On the other hand, the LA provides a set of APIs to manage the real-time data lifecycle for continuous learning. Moreover, the LA can process the live data to provide complex analysis creating real-time results for alerting or informing about important conditions in the factory, that may be not be seen at first glance. Finally, the LA allows the possibility to adapt to the productions needs during the production process.

The Complex-Event Machine Learning (CEML) [2] is a framework that combines Complex-Event Processing (CEP) [3] and Machine Learning (ML) [4] applied to the IoT. This means that the framework was developed to be deployed everywhere, from the edge of the network to the cloud. Furthermore, the framework can manage itself and works autonomously. The following section briefly describes the different aspects that CEML covers. The framework must automate the learning process and the deployment management. This process can be broken down in different phases: (1) the data must be collected from different sensors, either from the same device or in a local network. (2) The data must be pre-processed for attribute extraction. (3) The learning process takes place. (4) The learning must be evaluated. (5) When the evaluation shows that the model is ready, the deployment must take place. Finally, all these phases happen continuously and repetitively, while the environment constantly changes. Therefore, the model and the deployment must adapt as well.

6.3.3.1 Learning agents architecture

We utilize LinkSmart® LA following a modular architecture with loosely coupled modules responsible for different tasks. Figure 6.9 illustrates the architecture of the LA. The data and commands come via communication protocols implemented by Connectors (Figure 6.9 shows two example implementations, REST and MQTT). The connectors transfer the information to the Feeders, which process the data accordingly to the API logic. This logic depends on whether it is an insertion of new raw data, request of simple data processing (statement) or a machine learning request (CEML request). The data is inserted into the execution environment (in this case EsperEngine[1]), while the data processing requests are deployed in the same engine for the

[1] Esper is an open-source Java-based software product for Complex event processing (CEP) and Event stream processing (ESP) that analyzes series of events for deriving conclusions from them. See http://www.espertech.com/

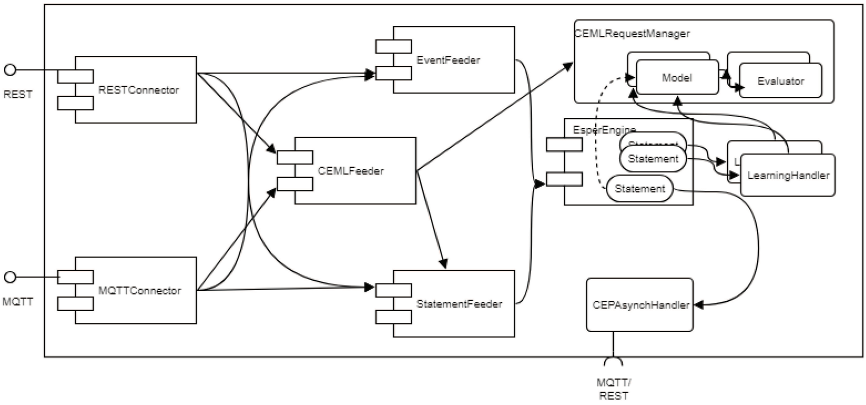

Figure 6.9 LinkSmart® Learning Service Architecture sketch.

processing of the raw data. The CEML request has a more complex behaviour. Each CEML request is managed by its own CEMLManager, which contains and coordinates the model(s), evaluator for each model, and several statements. Finally, all output of any process (Statement) in the execution pipeline (EsperEngine) is captured or managed by a Handler. If the process should be prepared and sent through a communication protocol, then it will be handled by a Complex-Event Handler: An Asynchronous Handler, if the protocol is asynchronous (e.g. MQTT); or Synchronous Handler, if the protocol is synchronous (e.g. HTTP).

6.3.3.2 Data propagation phase

Data in the IoT is produced in several places, protocols, formats, and devices. Although this article does not address the problem of data heterogeneity in detail, the learning agents require a mechanism to acquire and manage the heterogeneity of the data. The mechanism must be scalable and, at the same time, the protocol should handle the asynchronous nature of IoT. Finally, the protocol must provide tools to handle the pub/sub characteristics of the CEP engines. Therefore, we have chosen MQTT[2], a well-established Client Server publish/subscribe messaging transport protocol. The topic based message protocol provides a mechanism to manage the data heterogeneity by making a relation between topics and payloads. It allows deployments in several architectures, OS, and hardware platforms; basic constraints at the edge of the

[2] MQTT is a machine-to-machine (M2M)/“Internet of Things” connectivity protocol. Source http://mqtt.org/

network. The protocol is payload agnostic and as such allows for maximum flexibility to support several types of payloads.

6.3.3.3 Data pre-processing (munging) phase

Usually ML is tied to stored datasets, which incurs several drawbacks. Firstly, the learning can take place only with persistent data. Secondly, usually the models generated are based on historical data, not current data. Both constrains, in the IoT, have direct consequences. It is neither feasible nor profitable to store all data. In addition, embedded devices do not have much storage capacity, which makes it impossible to use ML algorithms on them. Furthermore, IoT deployments are commonly exposed to ever-changing environments.

Using historical data for off-line learning could cause outdated models to learn old patterns rather than current ones, producing drifted models. Although some IoT platforms like COMPOSITION support storage of historical data, it may be too time and space consuming to create large enough times series. Therefore, there is also a need for non-persistence manipulation tools. This is precisely what the CEP engine provides in the CEML framework. This means, the CEP engine decides which data and how the data is manipulated using predefined CEP statements deployed in the engine. Each statement can be seen as a topic, to which each learning model is subscribed. Any update of the subscribers provides a sample to be learnt in the learning phase.

6.3.3.4 Learning phase

There is no pre-selection of algorithms in the framework. They are selected by the restrictions imposed by the problem domain. For example, in extreme constrained devices, algorithms such as Algorithm Output Granularity (AOG) [5] may be the right choice. In other cases where the model changes quickly, one-shot algorithms may be the best fit. Artificial Neural Networks are good for complex problems but only with stable phenomena. This means that the algorithm selection should be made case-by-case. Our framework provides mechanisms for the management and deployment of the learning models, and the process of how the model is fed with samples. In general, the process is based on incremental learning [6] albeit with online and non-persistent data. The process can be summarized as follows: the samples, without the target provided in the last phase, are used to generate a prediction. The prediction will then be sent to the next phase. Thereafter, the sample is applied to update the model. Thus, all updates are used for the learning process.

6.3.3.5 Continuous validation phase

This section describes how the validation of the learning models is done inside the CEML. This phase does not influence the learning process nor validate the CEML framework itself.

ML model validation is a challenging topic in real-time environments and the evaluation for distributed environments or embedded devices is not addressed extensively in the literature, which is why we think it needs further research. There are two addressed strategies. Either we holdout an evaluation dataset by taking a control subset for given time-frame (time window), or we use *Predictive Sequential*, also known as *Prequential* [7], in which we assess each sequential prediction against the observation. The following section describes the continuous validation we applied for a **classification** problem, even though it can be applied for other cases as well.

Instead of accumulating a sample for validation, we analyse the predictions made before the learning takes place. All predictions are assessed each time an update arrives. The assessment is an entry for the confusion matrix [8], which is accumulated in an *accumulated confusion matrix*. The matrix contains the accumulation of all assessed predictions done before. In other words, the matrix does not describe the current validation state of the model, but instead the trajectory of it. Using this matrix, the accumulated validation metrics (e.g. Accuracy, Precision, Sensitivity, etc.) are being calculated. This methodology does have some drawbacks and advantages, explained more extensively in [9].

6.3.3.6 Deployment phase

The continuous validation opens the possibility for making an assessment of the status of the model each time a new update arrives, e.g. if it is accrued or not. Using this information, the CEML framework has the capability to decide if the model should or should not be deployed into the system at any time. If the model is behaving well, then it should be deployed, otherwise it should be removed from the deployment. The decision is made by user-provided thresholds w.r.t. evaluation metrics. If a threshold is reached, the CEML inserts the model into the CEP engine and starts processing the streams using the model. Otherwise, if the model do not reach the threshold, it is removed from the CEP engine.

6.3.3.7 Assessment

In [6] 13 issues for learning in the IoT where left open. The CEML framework addresses 10 out of the 13 challenges as follows:

- Handling the continuous flow of data streams: This is done by the stream statements inside the CEP engine using continuous streams for learning an evaluating.
- Unbounded memory requirements: The use of CEP engines in stream windows allows the intelligent usage of the memory as is needed, dropping it otherwise.
- Transferring data mining results over a wireless network with limited bandwidth: This is partially handled. MQTT is a reliable low-bandwidth lightweight protocol developed for satellite monitoring of pipelines. Nevertheless, this paper does not address the physical layer.
- Modelling changes of mining results over time: The CEML is a continuous automatic learning mechanism. The learning models will adjust as they learn.
- Interactive mining environment to satisfy user requirements: The IoT Learning agent provide an REST API. Thus, update the learning request is possible, as well as, obtaining live or on-demand updates.
- Integration between data-stream management systems and ubiquitous data-stream mining approaches: The CEML provides a REST API for managing each kind of request independently. Thus, the learning request can be managed as a whole, including the involved streams. Besides, the streams can be managed individually as single stream statement. Additionally, the MQTT API provides a multi-cast API so that in distributed multi-agent deployment, the agents can be managed as one, as groups, or as one entity.
- The relationship between the proposed techniques and the needs of real-world applications: Legal, ethical and technical reasons are part of the motivation. E.g. the storage constrains or the legal constraints in the health domain.
- Data pre-processing in the stream-mining process: This is handled in the pre-processing phase of the CEML.
- The technological issue of mining data streams: The implementation presented here shows that the system behaves in a real-time environment.
- The formalization of real-time accuracy evaluation: This is addressed by the Double-Tumble-Window Evaluation.

In addition to the Complex-Event Machine Learning approach based on the open-source IoT platform LinkSmart, we also describe another approach carried out in the scope of the project COMPOSITION in the next section.

6.3.4 Unsupervised Anomaly Detection in Production Lines

In addition to the previously introduced framework, which primarily allows for an exploitation of supervised machine learning algorithms, this chapter focuses on an alternative unsupervised approach that was also implemented in the scope of the project COMPOSITION. This method was used as a further extension to optimize the detection of machine errors in production lines at early stages.

In the last couple of years, the importance of cyber-physical systems in order to optimize industry processes, has led to a significant increase of sensorized production environments. Data collected in this context allows for new intelligent solutions to e.g. support decision processes or to enable predictive maintenance.

One problem related to the latter case is the detection of anomalies in the behaviour of machines without any kind of predefined ground truth. This fact is further complicated, if a reconfiguration of machine parameters is done on-the-fly, due to varying requirements of multiple items processed by the same production line. As a consequence, a change of adjustable parameters in most cases directly leads to divergent measurements, even though those observations should not be regarded as anomalies.

In the scope of the project COMPOSITION, the task of detecting anomalies for predictive maintenance within historical sensor data from a real reflow oven was investigated. While the oven is used for soldering surface mount electronic components to printed circuit boards based on continuously changing recipes, one related problem was the unsupervised recognition of potential misbehaviours of the oven resulting from erroneous components. The utilized data set comprises information about the heat and power consumption of individual fans. Apart from additional machine parameters like a predefined heat value for each section of the oven, it contains time-annotated sensor observations and process information recorded over a period of more than seven years.

As one solution for this problem, we will present our approach named Generic Anomaly Detection for Production Lines, short GADPL. The hereafter-presented description of GADPL is based on the stage-wise implementation of the algorithm. After an initial clustering of similar input parameters and a consecutive segmentation, we will discuss the representation of individual segments and the corresponding measurement of dissimilarity.

6.3.4.1 Configuration clustering

In many companies, as well as in the case of the project COMPOSITION, a single production line is often used to produce multiple items according to different requirements. Those requirements are in general defined by varying machine configurations consisting of one or more adjustable parameters, which are changed 'on-the-fly' during runtime. For a detection of deviations with respect to some default behaviour of a machine, this fact raises the problem of invalid comparisons between sensor measurements of dissimilar configurations. If a measurement or an interval of measurements is identified as an anomaly, it should only be considered as such, if this observation is related to the same configuration as observations representing the default behaviour. Therefore in advance to all subsequent steps, at first all sensor measurements have to be clustered according to their associated configuration. For the sake of simplicity, we are only discussing the process within a single cluster in the following subsections, although one has to keep in mind that each step is done in parallel for all clusters.

6.3.4.2 Segmentation

As a result of the configuration-based clustering, the data is already segmented coarsely. However, since this approach describes unsupervised anomaly detection, the idea of a further segmentation is to create some kind of ground truth, which reflects the default behaviour of a machine. In this section, we will see how the segmentation is utilized to implement this idea. In an initial step, a maximum segmentation length is defined, in order to specify the time horizon, after which an anomaly can be detected. Assuming a sampling rate of 5 mins per sensor, the maximum length of a segment would consequently be $(60 \times 24)/5 = 288$ to describe the behaviour on a daily basis. Although a decrease of the segment length implies a decrease of response time, it also increases the computational complexity and makes the detection more sensitive to invalid sensor measurements. In this context, it needs to be mentioned that in this stage segments are also spitted, if they are not continuous with respect to time as a result of missing values. Another fact that has to be considered is the transition time of configuration changes. While the input parameters associated with a configuration change directly, the observations might adapt more slowly and therefore blur the expressiveness of the new segment. To prevent this from happening, the transition part of all segments, which have been created due to configuration changes, is truncated. If segments become smaller than a predefined threshold, they can be ignored in the upcoming phases.

6.3.4.3 Feature extraction

Having a set of segments for each configuration, the next step is to determine the characteristics of all segments. While the literature presents multiple approaches to describe the behaviour of time series, we will focus on common statistical features extracted from each segment. Nonetheless, the choice of features is not fixed, which is why any feature suitable for the individual application scenario can be used. One example for rather complex features could be the result of a kernel fitting in the context of Gaussian processes, accepting a decrease in performance. Since the goal is to capture comparable characteristics of a segment, we compute different real-valued features and combine them in a vectorised representation. In the case of the project COMPOSITION, we used the mean to describe the average level, the variance as a measure of fluctuation and the lower and upper quartiles as a coarse distribution-binning of values. Due to the expressiveness of features being dependent from the actual data, one possible way to optimize the selection of features is the Principal Component Analysis. Simply using a large number of features to best possibly cover the variety of characteristics might have a negative influence on the measurement of dissimilarity. The reason for this is the partial consideration of irrelevant features within distance computations. Moreover, since thresholds could be regarded as a more intuitive solution compared to additionally extracted features, this replacement would lead to a significant decrease in the number of recognized anomalies. Apart from the sensitivity to outliers, the reason is a neglect of the inherent behaviour of a time series. As an example, consider the measurements of an acoustic sensor attached to a motor that recently is sending fluctuating measurements, yet within the predefined tolerance. Although the recorded values are still considered as valid, the fluctuation with respect to the volume could already indicate a nearly defect motor. Finally, one initially needs to evaluate appropriate thresholds for any parameter of each configuration.

6.3.4.4 Dissimilarity measurement

So far, we have discussed the exploitation of inherent information, extracted from segmented time series. The final step of GADPL is to measure the level of dissimilarity for all obtained representatives. Since no ground truth is available to define the default behaviour for a specific configuration, the algorithm uses an approximation based on the given data. One problem in this regard is the variability of a default behaviour, consisting of more than one pattern. Therefore, a naive approach as choosing the most occurring representative, would already fail for a time series consisting of two equally appearing patterns captured by different segments, where consequently half of the data would be detected as anomalous behaviour.

As one potential solution GADPL instead uses the mean over a specified size of nearest neighbours, depicting the most similar behaviour according to each segment. The idea is that even though there might multiple distinct characteristics in the data, at least a predefined number of elements represent the same behaviour compared to the processed item. Otherwise, this item will even have a high average dissimilarity with respect to the most similar observations and can therefore be classified as anomaly.

Here, for the vectorised feature representations, any suitable distance function is applicable. In the context of the project COMPOSITION we decided to use the Euclidean distance for a uniform distribution of weights, applied to normalized feature values. To further increase the performance of nearest neighbour queries, we exploited the R*-tree as a high-dimensional index structure. Given the dissimilarity for each individual representative together with a predefined anomaly threshold, GADPL finally emits potential candidates having an anomalous behaviour.

The application of GADPL is illustrated in Figure 6.10. The upper part shows the segmentation of time annotated power consumption data in percent. The lower part illustrates the result of the dissimilarity measurement, where the red rectangle indicates classified anomalies.

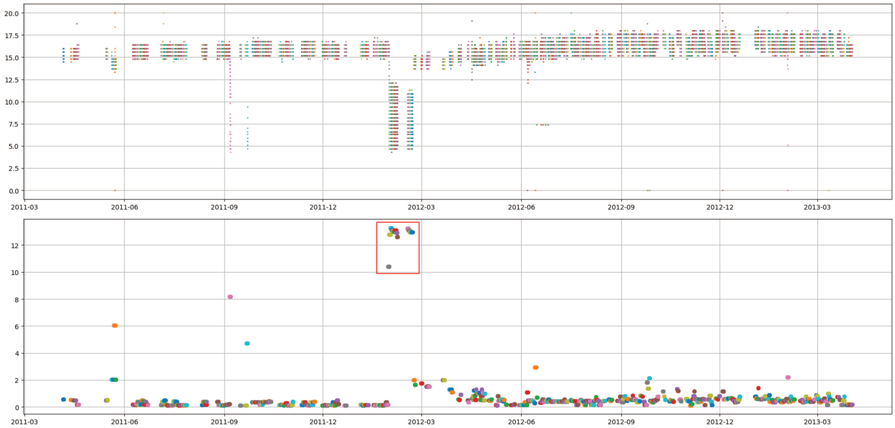

Figure 6.10 Example application of GADPL.

6.3.5 Summary

In this section, we have presented our recent research activities within the scope of the EU project COMPOSITION. As an example for the manufacturing industry, we have briefly described the COMPOSITION architecture along with one of its main components: the open-source IoT platform Link-Smart. As part of our ongoing work, we have also described our corresponding research activities regarding the analysis of sensor data from manufacturing industry.

6.4 Summary and Conclusions

In this work, we have given insights into our recent research activities with regard to the domains of Smart Data and Industrial Internet of Things. To this end, we have focused on the EU projects MONSOON and COMPOSITION as examples for the Public-Private Partnership (PPP) initiatives Factories of the Future (FoF) and Sustainable Process Industry (SPIRE). We have shown two different but conceptually similar architectures for scalable and agile data analytics. In addition, we have provided an overview of our recent Smart Data activities and have exemplified ongoing data-driven analysis of industrial production processes from the process and manufacturing industries.

We conclude that data-driven investigations, either applied in process industry or manufacturing industry, require a solid platform for handling data analytics at scale. The proposed architecture of the Cross Sectorial Data Lab in combination with the open-source IoT platform LinkSmart seem to be promising developments, which are applicable to any industrial sector.

Acknowledgements

The projects underlying this work have received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 723650 (MONSOON) and grant agreement No 723145 (COMPOSITION). This work reflects only the authors' views and the commission is not responsible for any use that may be made of the information it contains.

References

[1] C. Beecks, S. Devasya and R. Schlutter, "Data Mining and Industrial Internet of Things: An Example for Sensor-enabled Production Process Optimization from the Plastic Industry," *International Conference on Industrial Internet of Things and Smart Manufacturing (accepted),* 2018.

[2] J. Á. Carvajal Soto, M. Jentsch, D. Preuveneers und E. Ilie-Zudor, "CEML: Mixing and Moving Complex Event Processing and Machine Learning to the Edge of the Network for IoT Applications," *Proceedings of the 6th International Conference on the Internet of Things,* pp. 103–110, 2016.

[3] D. Bonino, J. A. Carvajal Soto, M. T. Delgado Alizo, A. Alapetite, T. Gilbert, M. Axling, H. Udsen und M. Spirito, "Almanac: Internet of things for smart cities," *2015 3rd IEEE International Conference on Future Internet of Things and Cloud (FiCloud),* pp. 309–316, 2015.

[4] G. Cugola und A. Margara, "Processing flows of information: From data stream to complex event processing," *ACM Computing Surveys (CSUR),* p. 15, 2012.

[5] C. Andrieu, N. De Freitas, A. Doucet und M. I. Jordan, "An introduction to MCMC for machine learning," *Machine learning,* pp. 5–43, 2003.

[6] M. M. Gaber, S. Krishnaswamy und A. Zaslavsky, "Advanced Methods for Knowledge Discovery from Complex Data," *On-board Mining of Data Streams in Sensor Networks,* pp. 307–335, 2005.

[7] N. A. Syed, S. Huan, L. Kah und K. Sung, "Incremental learning with support vector machines," *Citeseer,* 1999.

[8] A. P. Dawid, "Present position and potential developments: Some personal views: Statistical theory: The prequential approach," *Journal of the Royal Statistical Society. Series A (General),* pp. 278–292, 1984.

[9] G. J. Vachtsevanos, I. M. Dar, K. E. Newman und E. Sahinci, "Inspection system and method for bond detection and validation of surface mount devices." USA Patent 5,963,662, 1999.