Digital Shadows: Infrastructuring the Internet of Production

2

Wil van der Aalst, Matthias Jarke, István Koren, and Christoph Quix

Contents

Abstract

Digitization in the field of production is fragmented in very different domains, ranging from materials to production technology to process and business models. Each domain comes with specialized knowledge, often incorporated into mathematical models. This heterogeneity makes it hard to naively exploit advances

W. van der Aalst (✉) · I. Koren
Process and Data Science (PADS), RWTH Aachen University, Aachen, Germany
e-mail: wvdaalst@pads.rwth-aachen.de; koren@pads.rwth-aachen.de

M. Jarke
Information Systems and Databases (DBIS), RWTH Aachen University, Aachen, Germany

Fraunhofer Institute for Applied Information Technology (FIT), St. Augustin, Germany
e-mail: jarke@dbis.rwth-aachen.de

C. Quix
Fraunhofer Institute for Applied Information Technology (FIT), St. Augustin, Germany
e-mail: quix@fit.fraunhofer.de

C. Brecher et al. (eds.), *Internet of Production*, Interdisciplinary Excellence Accelerator Series, https://doi.org/10.1007/978-3-031-44497-5_25

in data-driven machine learning that could facilitate situation adaptation and experience transfer. Innovative combinations of model-driven and data-driven solutions must be invented but also made comparable and interoperable to avoid ending up in information silos. In future *World Wide Labs (WWLs)*, experiences can be shared, aggregated, and used for innovation. WWLs will be complex, evolving socio-technical networks of interconnected devices, software, data stores, and humans as users and contributors of expert knowledge and feedback. Integrating a large number of research labs, engineering, and production sites requires a capable cross-domain Internet of Production (IoP) infrastructure. The IoP project claims *Digital Shadows (DSs)* to offer a shared conceptual foundation for infrastructuring the IoP. In engineering, DSs were introduced as the data provision link to Digital Twins, whereas in computer science, DSs generalize the well-established concept of database views. In this chapter, we elaborate on the roles of DSs in infrastructuring the IoP from three perspectives: analytic functionality, conceptual organization, and technical networking. As an example where an integrative DS-like approach is already highly successful, we showcase the approach and infrastructure of the process mining field.

Keywords

Digital twin · Digital shadow · Data integration · Industry 4.0 · Internet of production · Manufacturing · Industrial infrastructure · Process mining

2.1 Introduction

Manufacturers need to handle vast volumes of heterogeneous, raw data with some machines capable of generating more than 1,000 different sensor signals, partly with enormously high sampling rates. Such amounts of data cannot be processed together close to their sources anymore. In addition, production processes may involve multiple machines, storage systems, transportation systems, and interactions with suppliers and logistic partners. Along industrial process chains, this complexity increases due to the different processes and, in some cases, within a single production line for reasons of variant diversity and mass customization.

Today, data is often stored without conceptual descriptions, engineering models, and their relationships, which prevents systematic reuse within and particularly across domains such as materials engineering, production technology engineering, operations, and management, as well as inter-organizational information exchange. The *Internet of Production (IoP)* project at RWTH Aachen University aims to investigate, prototypically demonstrate, and evaluate worldwide networks of production sites and research labs to cope with the challenges of industry like productivity, product variety due to make-to-order manufacturing, and sustainability. In our vision, a global interconnection of production sites and research labs forms the World Wide Lab (WWL), offering a controlled exchange of Digital Shadows even across organizational boundaries.

Compared to strategies like “Industry 4.0,” “Industrial Internet of Things,” and “Made in China 2025,” the IoP aims to go deeper in its cross-domain focus. It

requires novel domain-specific combinations of physical models and data-driven machine learning algorithms but also a common abstraction that makes them interoperable and exchangeable. The IoP project postulates the so-called Digital Shadows to be suitable for this task. As a situated (often real-time) means of reducing the statistical uncertainty of even the most advanced generic mathematical engineering models, DSs require the continuous integration of underlying data sources, which are heterogeneous in location, structure, and semantics, toward purpose-driven, aggregated, multi-perspective, and persistent datasets (Becker et al., 2021). The resulting datasets can feed and/or trace simulation models in Digital Twins. Our core hypothesis is that a well-designed collection of DSs is suitable as a cornerstone of infrastructures for designing, creating, and managing WWLs. This requires not only the powerful functionalities enabled by the novel construction of DS but also a suitable conceptual organization and physical infrastructure design.

According to theories of digital infrastructures (Pipek and Wulf, 2009), their successful creation and evolution require an iterative process of top-down design and bottom-up usage feedback, called infrastructuring. The IoP infrastructuring process, therefore, drives the further DS formalization and tooling with a stepwise buildup of more and more complex use cases, starting with local domain-specific and first data exchange experiments and prototypes, followed by more and more complex scenarios in WWLs among scientific and industrial partners.

This chapter is organized as follows. After reviewing recent related work in the fields of Digital Twins, Digital Shadows, and sovereign cross-organizational data sharing in Sect. 2.2, this chapter, therefore, discusses challenges for the functional requirements, conceptual modeling, and technical infrastructure of a DS-based IoP infrastructure in Sect. 2.3.

Complementing the already well-understood logic-based foundations and algorithms for heterogeneous data integration and analytics around structural data models (Jarke and Quix, 2017; Lenzerini, 2019), the DS-like formalisms of process mining constitute a highly successful core middle-ground abstraction for dynamics in the IoP, which we summarize in Sect. 2.4. Section 2.5 concludes the chapter.

2.2 Related Work on Digital Twins and Digital Shadows

Digital Twins have become a hot topic in the engineering literature in the last years, and several surveys have appeared. The DT concept was initially proposed by Grieves as a vision toward product life cycle management (Grieves, 2014). Their fundamental structure consists of a physical system and a corresponding computational model serving as its DT, which are dynamically synchronized through a mechanism known as *twinning*, cf. Fig. 2.1. The DT is generally regarded as a structural, optimization, or simulation model that represents the physical system. The twinning process involves two phases: the physical-to-virtual link, where physical system measurements are analyzed and the DT is modified accordingly, and the virtual-to-physical link, where the physical system is controlled using the information obtained from the DT.

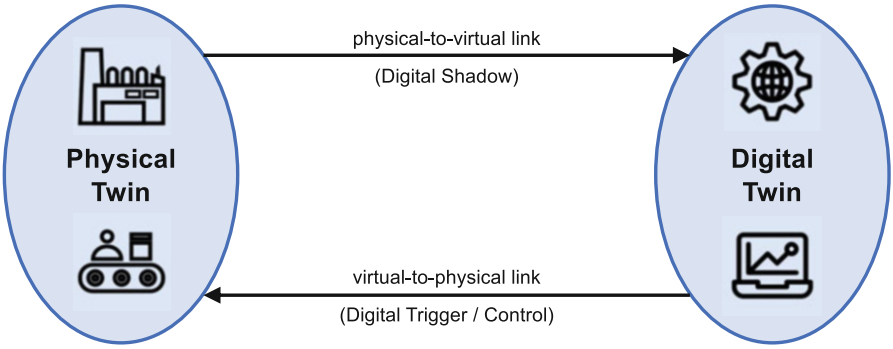

Fig. 2.1 Digital Twinning concept iterating between physical and Digital Twins (© the authors)

One significant development that has been consistently pursued in the context of the IoP is the recursive extension of the concept of a "physical system" to include cyber-physical (production) systems (CP(P)S). This extension involves managing federated networks of DT pairs within and across multiple interacting life cycles of engineering, production, and product usage. Lim et al. (2020) identify eight different perspectives and provide an in-depth analysis of engineering product life cycle management and business innovation. Meanwhile, Zhang et al. (2019) and Melesse et al. (2020) argue that research on DTs for product-service systems is becoming increasingly important due to the significant value that services can provide.

Fuller et al. (2020) identify Digital Shadows with the physical-to-virtual link, as shown in Fig. 2.1. However, the reality of this link is much more complex.

In their extensive review of the literature, Jones et al. (2020) divide the physical-to-virtual link into two parts: a *metrology* component that involves specifying and executing necessary measurements for real-time data analytics and a *realization* component that determines the changes required in the DT. Consistent with our findings in Pennekamp et al. (2019), they also highlight the IT requirements for implementing the link. These include advanced network algorithms based on Industrial IoT frameworks, such as those used for sensor and actuator management; efficient and effective information logistics over these networks; and data management, monitoring, and learning algorithms necessary for each step.

Bauernhansl et al. (2018) suggest a roadmap for examining the intricate information logistics that are necessary for DSs to efficiently and effectively provide information in today's dynamic industrial environment. Although a few studies have acknowledged the IoP's emphasis on the extensive integration of rapid mathematical engineering models and advanced data-to-knowledge pipelines using layers and networks of DSs as reusable objects, we are not aware of any studies that have investigated this topic in comparable depth and breadth.

Bazaz et al. (2020) highlight how the absence of clear data ownership can exacerbate these challenges. Moreover, a significant concern, especially by highly specialized production enterprises, is the loss of sovereignty over the use of their confidential know-how in such information logistics settings by keystone-player-driven data platforms or the violation of data privacy laws. To address these

problems, the concepts of Industrial Data Spaces and alliance-driven platforms have been derived in broad empirical studies (Otto and Jarke, 2019) and elaborated into a comprehensive reference architecture (Otto et al., 2022) for alliance-driven data exchange ecosystems. The IoP infrastructure adopts this basic sharing approach but makes it more specific by choosing Digital Shadows as the unit of knowledge sharing in inter-organizational data exchange. The arguments for this choice include, from a business perspective, the added value created by advanced methods for producing the Digital Shadow and the provenance documentation associated with that. From a technical perspective, the relatively small size of DSs compared to the vast amount of underlying data reduces network contention and enables the distribution of computations over multiple levels of Digital Shadows.

2.3 Infrastructure Requirements and DS Perspectives

The IoP infrastructure for data processing, AI, networking, and smart human interfaces needs to integrate methodologies from different perspectives, as summarized in Fig. 2.2. The requirements for the *functional perspective* include a range of powerful techniques for a huge variety of short-, medium-, and long-term tasks but also for task layers all the way from basic data integration up to model-integrated AI shadows and suitable interactive visualizations. In designing all these, there should be a balance between general and still directly applicable data science and ML techniques, keeping the continued shortage of human specialists in mind.

For the *conceptual foundation* linking these many aspects of DS, data must have semantics but at the same time be generic. Much of the existing modeling work in this area is either too general that one cannot apply interesting data science/ML techniques or it is very specific for a particular application.

Distributed execution and controlled sharing of data analysis functions with high performance, reliability, and security is addressed from a *physical perspective*. This perspective aims to provide an interconnected technical backbone of the Internet of Production, as sketched in the lower layer of Fig. 2.2.

The following subsections elaborate on these challenges in some more detail.

2.3.1 Functional Perspective: Data-to-Knowledge Pipelines Using Domain-Specific Digital Shadows

Digital Shadows provide domain-specific access to heterogeneous data from different sources, structures, and semantics. They prepare the application of data-driven machine learning methods, embedding engineering knowledge in the form of physical, predictive, and simulation models to raise relevance, performance, and explainability. For example, reduced mathematical models can be exploited as evaluation functions in neural network-based learning; conversely, the statistical uncertainty handling of generic mathematical models can be made significantly simpler and more precise by continuous monitoring of actual situations. Generally speaking, this requires deep and broad extensions to the emerging field called informed machine learning, as surveyed by von Rueden et al. (2021).

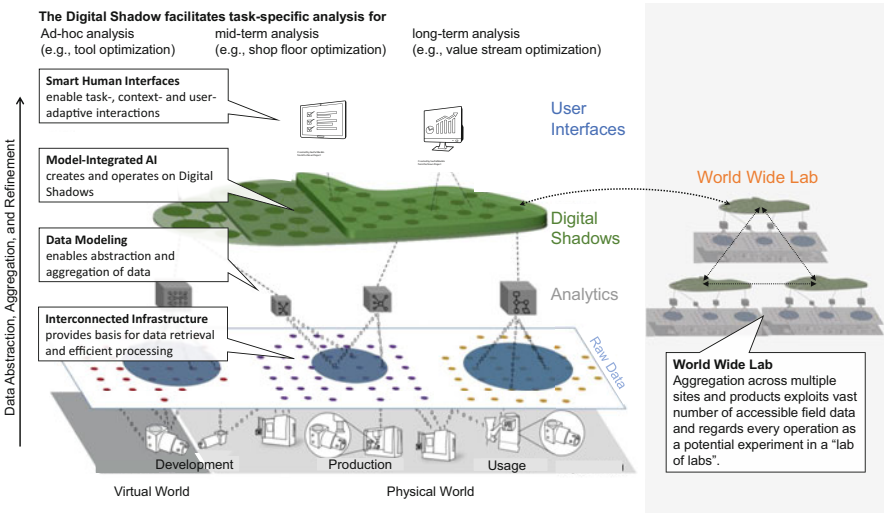

Fig. 2.2 Interconnected infrastructure for Digital Shadows in the Internet of Production (Brauner et al., 2022) (© the authors)

The data-to-knowledge pipelines (pathways from lower to higher levels in Fig. 2.2) help to transform massive data into insights while providing meaningful actionable knowledge to decision-makers.

Today, the transferability of learning outcomes between fields of application has only been realized in a few particular contexts. Correspondingly, the network of production-specific data-to-knowledge pipelines could enable the derivation of similarities between pipelines across different production settings to enable transfer learning within and across domain or organizational boundaries. As one example, the IoP project developed a user-centered planning tool with an integrated decision support system based on human-centered AI (Schemmer et al., 2020) that demonstrates how to increase the efficiency and reproducibility of planning process chains for fiber-reinforced plastics production.

Due to the massive sensor stream processing and real-time analytics in the IoP, data cleaning and integration may only happen on demand. The current numerical design of manufacturing processes based on Digital Twin simulations is unsuitable to directly support real-time decision-making at the machine. The Digital Shadow is based on a reduced simulation model, which focuses on only currently relevant aspects. This data must therefore be adequately provided and composed.

Since this data-driven approach blurs the distinction between design time and runtime of production systems, novel validation techniques need to be developed to account for and monitor adaptations in functionality, contexts, and constraints. Using AI methods allows distributed planning and scheduling as well as execution of production processes at all levels. Novel visualization and interaction methods need to be developed that minimize biased decision-making and support the

understanding of complex data. Process mining can be used to diagnose and improve quality and performance problems in complex systems with high concurrency.

2.3.2 Conceptual Perspective: Organizing DS Collections in a WWL

Although the infrastructure for the IoP should be *generic*, Digital Shadows need to be contextualized using purpose-oriented metamodels and ontologies. Data-driven approaches (e.g., machine learning, visual analytics, heterogeneous data integration, or process mining) require data and models in particular formats. On the one hand, aspects such as latency, scalability, connectivity, and security can be largely addressed while abstracting from the semantics of data and models. On the other hand, data can only be used in a meaningful way if there are suitable metamodels and/or ontologies. Models also need a specific structure and semantics in order to be used in an expressive manner.

For example, one can view both Petri nets and neural networks as graphs. However, this is not very meaningful. Analysis techniques for both classes of models are disjoint. Hence, data and models need to have structure and semantics to be meaningful in the context of the IoP. However, one should also avoid building an infrastructure and using data formats and models that are specific for a particular application (e.g., data generated by a particular machine). Different applications should use the same IoP infrastructure, and data formats and models should be reusable. One does not want to create new storage formats or machine learning techniques when a new machine is added. In other words, we need to find a trade-off between keeping things as generic as possible and at the same time being specific enough to create standardized storage formats and model types that enable the creation of data-driven techniques that help in decision-making and process improvement.

It is essential that the huge amount of data is organized and contextualized according to metamodels. The metamodels must be integrated into a cross-disciplinary life cycle with iterative data aggregation along sequential engineering stages. This requires concepts on storing and linking data from distributed stores within the Digital Shadow.

Specifically, partial metamodels are being investigated from the following perspectives:

- The bottom-to-top pipeline in Fig. 2.2 derives a hierarchy of DSs using principles of database view management in heterogeneous data integration and mining (Liebenberg and Jarke, 2020).
- From a software engineering and research data management perspective, DSs are software artifacts created by model-driven generation and documented according to FAIR principles with full relevant provenance and context information (Becker et al., 2021).
- From an analytics and specifically process mining perspective, data must be interpreted and integrated under an event-centric metamodel, cf. Sect. 2.4.

- From a cross-organizational sharing perspective, DSs are valuable and therefore threatened exchange objects for which a metamodel must allow the representation and monitoring of suitable service-oriented policies and business models (Jarke, 2020).
- A metamodel of the physical infrastructure underlies the secure and reconfigurable workflows of efficient, safe, and secure distributed computation, storage, and transport in complex physical networks (cf. Sect. 2.3.3).

Following established practice in metamodeling for method engineering (Jeusfeld et al., 2010), each of these perspectives needs individual "middle-ground" abstractions in the form of dedicated reference metamodels (see example in Sect. 2.4) whose inter-relationships can be maintained by their linkage to a generic meta-metamodel of DSs with shared domain terms. Such a well-organized collection of interrelated metamodels is under iterative development in the IoP cluster, reflecting ongoing experiences with many specific use cases within and beyond the cluster.

2.3.3 Physical Perspective: Interconnected Technical Infrastructure

The envisioned Internet of Production infrastructure ranges from monitoring and control information at the shop level to process development and analysis. Achieving this requires a combination of network infrastructure measures and scalable data stream processing techniques, along with decentralized process control methods, to enable high-performance, reliable, safe, and secure distributed communication networks that support distributed multi-agent model executions and data flows. A dynamically reconfigurable architecture for these production-specific data flows then enables secure industrial cooperation, which in turn leads to a steep increase in data produced and consumed.

The requirements of the technical infrastructure can be grouped according to three core challenges: *Seamless low latency* enables adaptive control operations within network infrastructures. *High-performance* adaptive stream processing components provide scalability. *Security* is key in industrial cooperation scenarios through data security, data sovereignty, and stakeholder confidentiality.

The specific trait of these challenges is a trade-off, depending on whether data is in motion, in use, or at rest. For example, for scalable data processing, data are moved across a network to a cloud environment, thereby improving performance while increasing latency and possibly decreasing security. Data in use resides in non-volatile media where it can be processed with low latency. Data at rest, stored in data warehouses or data lakes, enables long-term observation and analysis.

Starting with near-to-machine edge processing, the technical infrastructure continues with processing rules that can be efficiently implemented in hardware in a WWL. For example, an early IoP demonstrator (Pennekamp et al., 2020) demonstrates executing performance comparisons among companies in the injection molding industry under encryption, i.e., without revealing companies' sensitive

data to anyone. The approach utilizes homomorphic encryption to protect sensitive information when performing computations on joint data.

2.3.4 Toward an Empirically Grounded IoP Infrastructure

A core goal of the IoP project is to find an overall architectural approach that brings these perspectives together. Toward this purpose, numerous individual use cases and experiments are conducted concerning different domains and perspectives of DS development and usage, as reported in the remaining chapters of this book.

A series of increasingly powerful partial operational infrastructures are needed to do this. On the one hand, they must allow researchers to demonstrate and evaluate the DS-based approach of the cluster from engineering, social, economic, and IT perspectives. However, on the other hand, they must interoperate with commercially used tools to enable cooperation with existing lab equipment and industrial environments. AI-inspired multi-agent architectures have been studied in Liebenberg (2021) as a promising approach to bring several of the mentioned DS perspectives together. It partially automates the search for data, knowledge, and Digital Shadows and demonstrates that a combination of social and technical agents is feasible in a WWL yet requires semantic interoperability to achieve the needed provenance and explainability.

The infrastructuring approach maps formal strategic dependency and goal models down to software agents executed in a Kubernetes infrastructure that is in turn linking diverse professional data management systems (loosely integrated as a data lake) and newly developed microservices. The technical IT infrastructure was built on an open-source software stack ensuring interoperability to commercial tools such as Azure and MindSphere. Technically, this lays the foundation for automated data streaming of sensor data from machines to their analysis by dynamically providing connectivity, storage, and computing resources. The embedded data lake permits the handling of relational, graph-based, document-oriented, and time-series data.

To illustrate the combination of IoP-specific and existing commercial resp. open-source technologies in this multi-agent architecture, we briefly sketch its application in a quite demanding engineering use case.

The high-pressure die casting process is a highly automated production technology that generates large amounts of data. Yet, the extensive breadth (number of values) and depth (frequency and precision) require domain knowledge of the process to select required data to facilitate product quality and productivity improvements. In an experimental setup, Rudack et al. (2022) and Chakrabarti et al. (2021) accessed multiple data sources based on the Open Platform Communication Unified Architecture (OPC-UA) and transmitted them to a streaming pipeline defined in the low-code programming environment Node-RED and executed by Apache Kafka. The data are subsequently stored in our data lake, utilizing a MinIO object store as the underlying storage system. A systematic hierarchy of high-dimensional DSs for data analytics is combined with an AI-based recommender system for interactive visual analytics, e.g., in product quality assurance (Chakrabarti et al., 2021).

Tests confirm decent analytics capabilities and interactive usability at the functional level, which deeper semantic models of the process will further improve. They also show an industrially relevant performance for the use case, which requires up to a few hundred messages per second.

2.4 Example of a Successful DS-Based Metamodel: Process Mining

As stated before, for the most relevant perspectives on the IoP, we need to find trade-offs between keeping things as generic as possible and being specific enough to create standardized storage formats and model types that enable the creation of data-driven techniques that help in decision-making and process improvement. In this section, we use *process mining* as an example technology that illustrates such a trade-off nicely. Process mining is generic and not tailored toward specific processes but provides a range of supporting techniques and tools. Moreover, process mining is well-suited to analyze and improve production processes. For example, most car manufacturers (e.g., BMW, Volkswagen, Ford, Toyota, Skoda, Fiat, Porsche, and Ferrari) already use process mining, e.g., to ensure the timely delivery of parts from suppliers, to optimize painting and assembly processes, to distribute cars, and to improve maintenance. In the context of IoP, we analyzed, for example, the production of e.GO cars at the plant in Aachen.

The starting point for process mining is *event data*. An event refers to an *activity* happening at a particular point in *time*. In a classical event log, each event refers to precisely one *case*. An event may have many more attributes (e.g., cost, resource, location, and organizational unit). However, the attributes activity, timestamp, and case are mandatory. Process mining aims to improve operational processes by systematically using such event data (van der Aalst, 2016). Process mining techniques utilize a combination of event data and process models to gain insights, identify bottlenecks and deviations, anticipate and diagnose performance and compliance issues, and facilitate the automation or elimination of repetitive tasks (van der Aalst and Carmona, 2022). The process mining discipline focuses on concrete tasks such as process discovery (turning event data into process models (van der Aalst, 2016)) and conformance checking (diagnosing differences between modeled and observed behavior (van der Aalst, 2016; Carmona et al., 2018)).

There are several open-source process mining tools; the best-known are ProM, PM4Py, RapidProM, and BupaR. There are also over 40 commercial process mining tools (see processmining.org for an overview). It is estimated that already over half of the Fortune 500 are applying process mining (Reinkemeyer, 2020). Examples include Deutsche Bahn, Lufthansa, Airbus, ABB, Siemens, Bosch, AkzoNobel, Bayer, Neste, Pfizer, AstraZeneca, MediaMarkt, Zalando, Uniper, Chevron, Shell, BP, Dell, Nokia, and the car manufacturers mentioned before.

There is a growing consensus that the assumption that each event refers to precisely one case is limiting. This is particularly relevant when analyzing pro-

duction processes. One assembly step may involve different parts, a machine, and an operator. This leads to the well-known convergence and divergence problems (van der Aalst and Berti, 2020; van der Aalst, 2021a). The *convergence problem* surfaces when a fine-grained case notion is used and the flattening of the event data leads to the unintentional replication of events, e.g., an assembly step is replicated for all the parts involved in it. The *divergence problem* appears when a course-grained case notion is used and causal relations between events get lost. Classical process mining forces the adoption of a single view of the processes under consideration. *Object-centric process mining (OCPM)* addresses the limitation by allowing for any number of objects per event (van der Aalst and Berti, 2020; van der Aalst, 2021a). This extension is highly relevant for the IoP and its Digital Shadows. Examples of objects are products, sub-assemblies, parts, robots, workers, machines, conveyor belts, etc. Events correspond to transformation, transportation, and assembly steps. In IoP, we analyzed assembly processes using real-world data from Heidelberger Druckmaschinen AG, a global manufacturer of printing presses. Heidelberger's printing presses are composed of many different parts, making the traditional case notion limiting. Therefore, we combined object-centric process mining where objects are organized in bills-of-material (Brockhoff et al., 2022).

OCPM uses the *metamodel* shown in Fig. 2.3. The metamodel is generic but specific enough to allow for analysis, discovery, conformance checking, decision-making, and automated process improvements. As mentioned earlier, the IoP strongly relies on data that have clear semantics and that allow for a range of

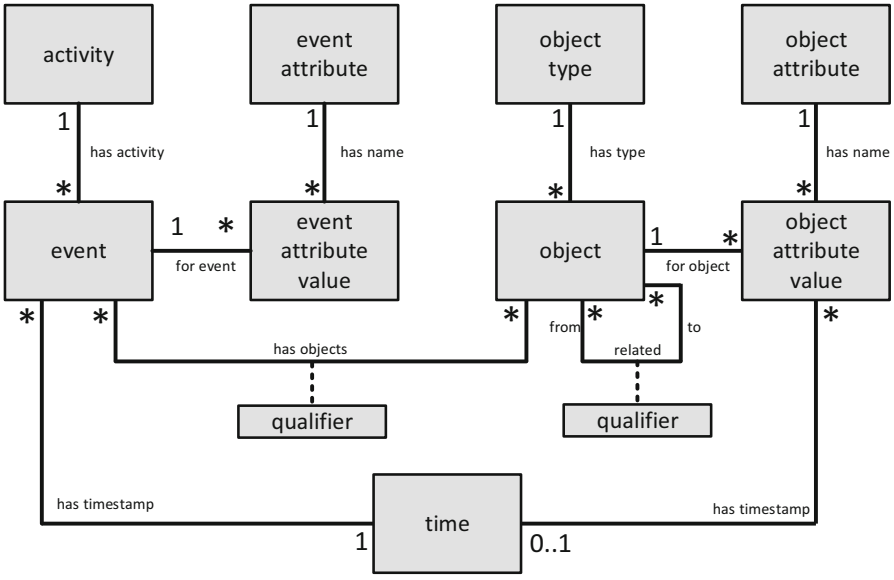

Fig. 2.3 Process mining metamodel: events have a timestamp, an activity, and other event attributes. An event may refer to any number of objects. Each object has a type and attributes that may change over time

techniques, without being application-specific. The two main ingredients of the metamodel shown in Fig. 2.3 are *events* and *objects*. An event has one activity and a timestamp. In addition, an event may have additional attributes such as costs, energy, coordinates, etc., and an event may refer to any number of objects. Each object has a type and may have any number of additional attributes. The values of object attributes may change over time. Therefore, object attribute values may have a timestamp. Object attribute values without timestamps can be seen as immutable. Object attribute values with timestamps can be seen as updates ordered in time. Objects may refer to other objects using a qualified relationship. For example, it is possible to state that one object is part of another object. There is a many-to-many relationship between events and objects. Also, this relationship is qualified. One may say that an event uses a set of objects or creates a set of objects. Note that when using traditional process mining, there is just one type of objects (called cases) and each event refers to precisely one such object. As mentioned, in IoP, we would like to consider many different types of objects concurrently, including machines, workers, orders, end-products, parts, organizational units, locations, shipments, suppliers, etc.

Figure 2.3 shows an example of a concrete metamodel giving meaning to events, objects, relationships, and attributes. The *object-centric event log (OCEL)* standard provides a storage and exchange format for such object-centric event data (Ghahfarokhi et al., 2021). Note that OCEL makes a few simplifying assumptions, e.g., attribute values cannot change and relationships are not qualified. However, both OCEL and Fig. 2.3 agree on the core concepts. Using such object-centric event data, it is possible to discover object-centric process models and check conformance automatically (van der Aalst and Berti, 2020; van der Aalst, 2021a). For example, van der Aalst and Berti (2020) shows that it is possible to automatically learn object-centric Petri nets showing frequencies, delay distributions, and probabilities in an integrated process model describing the interactions between any number of object types. Next to process discovery and conformance checking, there exist techniques to predict the behavior of object-centric processes, detect concept drift, analyze performance, and recommend actions to reduce operational friction.

By selecting a set of object types and a set of activities, one can easily create views on the whole. For example, one can focus on particular machines, products, and phases of the production process. For each view, it is possible to automatically create process models showing performance and compliance problems.

Such views are composed of object-centric event data projected onto selected object types and activities and object-centric process models. These provide concrete Digital Shadows that can be used to manage and improve production processes. The metamodel shown in Fig. 2.3 strikes a balance between generality and specificity. Application- or domain-specific data need to be mapped onto generic concepts such as events and objects, thus allowing for a range of techniques implemented in existing process mining tools.

Figure 2.4 shows the process mining pipeline (van der Aalst, 2021a). The first step is to extract *object-centric event data* from existing data sources (e.g., ERP and CRM systems). Such data can be *preprocessed*, e.g., selecting activities and object

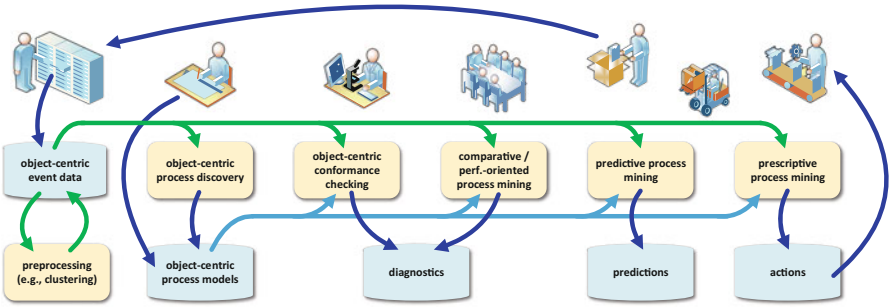

Fig. 2.4 The process mining pipeline showing the different types of artifacts and data processing steps starting from object-centric event data

types, or using automated clustering. For the resulting views, dedicated object-centric process models can be derived, thus forming Digital Shadows. Object-centric *process discovery* techniques can derive object-centric Petri nets, Business Process Model and Notation (BPMN), and Directly-Follows Graph (DFG) models without any modeling.

Given the object-centric event data and object-centric process models, one can apply object-centric *conformance checking* to find and diagnose deviations. Similarly, one can apply *comparative and performance-oriented process mining* techniques to diagnose execution gaps, i.e., significant differences between best practices and actual process executions (van der Aalst et al., 2021).

In IoP, we develop open-source software tools such as OCPM (www.ocpm.info) and OCPI (www.ocpi.ai) to support object-centric process mining using OCEL (Ghahfarokhi et al., 2021). However, our ideas have also been implemented in commercial software systems. A notable example is ProcessSphere by Celonis. This helps us to realize a World Wide Lab leveraging event data from different organizations. Process mining techniques can provide *backward-looking* or *forward-looking* analysis. Backward-looking analysis involves identifying the root causes of bottlenecks in production processes, while forward-looking analysis involves predicting the remaining processing time of ongoing cases and recommending actions to reduce failure rates. Both types of analysis can lead to actionable insights, such as implementing countermeasures to address performance or compliance issues (van der Aalst and Carmona, 2022). Figure 2.4 shows *predictive process mining* as an example of a forward-looking form of process mining. This results in predictions that can be used proactively. The final step in the pipeline depicted in Fig. 2.4 is *prescriptive process mining*. In this step, event data, process models, and objectives are combined to trigger actions addressing observed or predicted performance and compliance problems.

One main challenge is to extract event data from the source systems. Event data may exist at different levels of granularity, and often there are data quality problems. Once the data is extracted, cleaned, and stored using the metamodel in Fig. 2.3, the whole pipeline depicted in Fig. 2.4 can be applied.

Another major challenge is the collection of event data across organizational boundaries. Sharing event data may not be possible for some organizations, and they may use unique identifiers and logging practices (van der Aalst, 2021b). *Federated process mining* aims to tackle these problems by creating cross-organizational event data in such a way that confidentiality is ensured (van der Aalst, 2021b). Federated event logs make it possible to compare processes in different organizations and to analyze processes spanning multiple organizations.

Process mining techniques benefit directly from a technical infrastructure that is scalable, reliable, and safe. Process mining provides a strong conceptual foundation for realizing Digital Shadows and a World Wide Lab. The metamodel in Fig. 2.3 shows that it is possible to provide generic technologies close to production processes. The wealth of process mining tools and techniques supports the functional perspective of the IoP infrastructure.

Although the scope of process mining is broad and covers all optional processes, it is just one building block of the bigger IoP infrastructure. For example, techniques for process mining do not support continuous processes and unstructured data (e.g., computer vision and object recognition).

2.5 Conclusion

The infrastructure of the Internet of Production (IoP) research cluster aims, in the long term, to significantly facilitate the design, operation, and usage of World Wide Labs for more effective, scalable, safe, and secure data and knowledge sharing and usage across boundaries of domain, organizations, and even cultures. To bridge these boundaries, this chapter presented Digital Shadows as a core metaphor across all perspectives of the IoP infrastructure.

We showed that DSs have many different facets and roles, each of them requiring specific theoretical foundations, such as supporting (meta)models, algorithms, and software tools. For some facets, such as process mining or heterogeneous data integration, existing foundations just need to be adapted to some special requirements of the production sector; for others, we still need to identify such "middle-ground abstractions" that make overly abstract meta-metamodels more usable while still providing practical improvements for individual use cases and customer applications.

These theoretical developments are empirically confronted with a large number of interdisciplinary IoP use cases across research labs and with practice partners. To enable these use cases, a series of increasingly powerful experimental infrastructures, including linkage to widely used existing systems, are being developed.

While the present chapter reviewed the overall vision, challenges, and an integrative DS example as an infrastructuring concept, the status achieved in the first three project years is presented in the following three chapters of this book, addressing technical details, initial research results, and use cases for the physical, conceptual, and functional-algorithmic perspectives.

Acknowledgments Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-2023 Internet of Production – 390621612.

References

van der Aalst W (2016) Process mining: data science in action. Springer, Berlin/Heidelberg

van der Aalst W (2021a) Concurrency and objects matter! Disentangling the fabric of real operational processes to create digital twins. In: Cerone A, Olveczky P (eds) International Colloquium on Theoretical Aspects of Computing (ICTAC 2021). Lecture Notes in Computer Science, vol 12819. Springer, Berlin, pp 3–17

van der Aalst W (2021b) Federated process mining: exploiting event data across organizational boundaries. In: Atukorala N, Chang C, Damiani E, Fu M, Spanoudakis G, Srivatsa M, Wang Z, Zhang J (eds) IEEE International Conference on Smart Data Services (SMDS 2021). IEEE, pp 1–7

van der Aalst W, Berti A (2020) Discovering object-centric petri nets. Fund Inform 175(1-4):1–40

van der Aalst W, Carmona J (eds) (2022) Process mining handbook, Lecture Notes in Business Information Processing, vol 448. Springer, Berlin

van der Aalst W, Brockhoff T, Farhang A, Pourbafrani M, Uysal M, van Zelst SJ (2021) Removing operational friction using process mining: challenges provided by the internet of production (IoP). In: Hammoudi S, Quix C (eds) Data management technologies and applications. Communications in Computer and Information Science, vol 1446. Springer, Berlin, pp 1–31

Bauernhansl T, Hartleif S, Felix T (2018) The digital shadow of production – a concept for the effective and efficient information supply in dynamic industrial environments. Proc CIRP 72:69–74

Bazaz SM, Lohtander M, Varis J (2020) Availability of manufacturing data resources in digital twin. Proc Manuf 51:1125–1131

Becker F, Bibow P, Dalibor M, Gannouni A, Hahn V, Hopmann C, Jarke M, Koren I, Kröger M, Lipp J, Maibaum J, Michael J, Rumpe B, Sapel P, Schäfer N, Schmitz GJ, Schuh G, Wortmann A (2021) A conceptual model for digital shadows in industry and its application. In: Ghose A, Horkoff J, Silva Souza VE, Parsons J, Evermann J (eds) Conceptual modeling, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp 271–281. https://doi.org/10.1007/978-3-030-89022-3_22

Brauner P, Dalibor M, Jarke M, Kunze I, Koren I, Lakemeyer G, Liebenberg M, Michael J, Pennekamp J, Quix C, Rumpe B, van der Aalst W, Wehrle K, Wortmann A, Ziefle M (2022) A computer science perspective on digital transformation in production. ACM Trans Internet Things 3(2):1–32. https://doi.org/10.1145/3502265

Brockhoff T, Uysal MS, Terrier I, Göhner H, van der Aalst WMP (2022) Analyzing multi-level BOM-structured event data. In: Munoz-Gama J, Lu X (eds) Process mining workshops, vol 433. Springer International Publishing, Cham, pp 47–59. https://doi.org/10.1007/978-3-030-98581-3_4, Series Title: Lecture Notes in Business Information Processing

Carmona J, van Dongen B, Solti A, Weidlich M (2018) Conformance checking: relating processes and models. Springer, Berlin

Chakrabarti A, Sukumar R, Jarke M, Rudack M, Buske P, Holly C (2021) Efficient modeling of digital shadows for production processes: a case study for quality prediction in high pressure die casting processes. In: 8th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2021, Porto. IEEE, pp 1–9

Fuller A, Fa Z, Day C, Barlow C (2020) Digital twin: enabling technologies, challenges and open research. IEEE Access 8:108952

Ghahfarokhi A, Park G, Berti A, van der Aalst W (2021) OCEL Standard. www.ocel-standard.org

Grieves M (2014) Digital twin – manufacturing excellence through virtual factory replication. In: White Paper, LLC

Jarke M (2020) Data sovereignty and the Internet of Production. In: Advanced Information Systems Engineering, 32nd International Conference; Grenoble. Springer, Cham, pp 549–558

Jarke M, Quix C (2017) On warehouses, lakes, and spaces: the changing role of conceptual modeling for data integration. In: Conceptual modeling perspectives. Springer, pp 231–245. https://doi.org/10.1007/978-3-319-67271-7_16

Jeusfeld M, Jarke M, Mylopoulos J (2010) Metamodeling for method engineering. MIT Press, Cambridge, Mass

Jones D, Snider C, Nassehi A, Yon J, Hicks B (2020) Characterizing the digital twin: a systematic literature review. CIRP J Manuf Sci Technol 29:36–52

Lenzerini M (2019) Direct and reverse rewriting in data interoperability. In: International Conference on Advanced Information Systems Engineering. Springer, Cham, pp 3–13

Liebenberg M (2021) Autonomous agents for the world wide lab. PhD thesis, RWTH Aachen University

Liebenberg M, Jarke M (2020) Information systems engineering with digital shadows: concept and case studies. In: International Conference on Advanced Information Systems Engineering. Springer, Cham, pp 70–84

Lim K, Zheng P, Chen CH (2020) A state-of-the-art survey of digital twin: techniques, engineering product lifecycle management and business innovation perspectives. J Intell Manuf 31:1313–1337

Melesse T, di Pasquale V, Riemme S (2020) Digital twin models in industrial operations: a systematic literature review. Proc Manuf 42:267–272

Otto B, Jarke M (2019) Designing a multi-sided data platform: findings from the international data spaces case. Electr Mark 29(4):561–580

Otto B, ten Hompel M, Wrobel S (2022) Designing data spaces – the ecosystem approach to competitive advantage. Springer, Cham

Pennekamp J, Glebke R, Henze M, Meisen T, Quix C, Hai R, Gleim L, Niemietz P, Rudack M, Knape S, Epple A, Trauth D, Vroomen U, Bergs T, Brecher C, Buhrig-Polaczek A, Jarke M, Wehrle K (2019) Towards an infrastructure enabling the internet of production. In: 2019 IEEE International Conference on Industrial Cyber Physical Systems (ICPS). IEEE, pp 31–37. https://doi.org/10.1109/ICPHYS.2019.8780276

Pennekamp J, Sapel P, Fink IB, Wagner S, Reuter S, Hopmann C, Wehrle K, Henze M (2020) Revisiting the privacy needs of real-world applicable company benchmarking. LUIS, Leibniz Universität IT Services, Hannover, pp 31–44. https://doi.org/10.18154/RWTH-2021-01321

Pipek V, Wulf V (2009) Infrastructuring: Toward an integrated perspective on the design and use of information technology. J AIS 10(5):447–473

Reinkemeyer L (2020) Process mining in action: principles, use cases and outlook. Springer, Cham

Rudack M, Rath M, Vroomen U, Bührig-Polaczek A (2022) Towards a data lake for high pressure die casting. Metals 12(2):349. https://doi.org/10.3390/met12020349

von Rueden L, Mayer S, Beckh Kea (2021) Informed machine learning – a taxonomy and survey of integrating prior knowledge into learning systems. IEEE Trans Knowl Data Eng 1–19. https://doi.org/10.1109/TKDE.2021.3079836

Schemmer T, Brauner P, Schaar AK, Ziefle M, Brillowski F (2020) User-centred design of a process-recommender system for fibre-reinforced polymer production. In: Yamamoto S, Mori H (eds) Human interface and the management of information. Interacting with information, vol 12185. Springer International Publishing, Cham, pp 111–127. https://doi.org/10.1007/978-3-030-50017-7_8

Zhang H, Ma L, Sun J, Lin H, Thuerer M (2019) Digital twin in services and industrial product service systems: review and analysis. Proc CIRP 83:57–60