A Smart Manufacturing Data Lake Metadata Framework for Process Mining

Michalis Pingos
Department of Electrical Engineering, Computer Engineering and Informatics
Cyprus University of Technology
Limassol, Cyprus
email: michalis.pingos@cut.ac.cy

Andreas S. Andreou
Department of Electrical Engineering, Computer Engineering and Informatics
Cyprus University of Technology
Limassol, Cyprus
email: andreas.andreou@cut.ac.cy

***Abstract*— The fourth industrial revolution consists of a new level of organization and control of the entire production process. The smart manufacturing ecosystem and especially Cyber Physical Systems are evolving rapidly. They constitute an environment with multiple heterogeneous sources that produce high volumes of data. This data needs to be stored in a storage system that can handle raw, unprocessed, relational, and non-relational data types, such as Data Lakes, in order to be processed when needed. This paper introduces a Data Lake-based metadata framework, which utilizes the concept of blueprints to characterize the data sources and the data itself to facilitate process mining tasks. The applicability and effectiveness of the proposed framework is validated through a real-world smart manufacturing case-study, namely a poultry meat production factory, which offers operational support and business workflow analysis.**

Keywords: Smart Manufacturing; Data Lakes; Heterogeneous Data Sources; Metadata Mechanism; Data Blueprints; Process Mining.

I. INTRODUCTION

Industry 4.0 is based on a number of new and innovative technological developments, such as Cyber Physical Systems (CPSs), Internet of Things (IoT), Cloud Computing, Cognitive Computing, robotics, Augmented Reality (AR) technology and intelligent tools [1], which contribute to the production of personalized products according to customer needs by digitization of the entire product production cycle [2].

Factories of the future will consist of a CPS or a set of CPSs that will interact with each other. A CPS consists of mechanisms controlled or monitored by computer algorithms, integrated into the Internet and its users. CPSs change the way people interact with machines and workers will need to be skilled and will need to be aware of the functions of coordinated intelligent machines from a central control point and of the data they produce [3].

The main challenges of Industry 4.0 are the need for: (i) constant availability of all data and information in a smart factory in real-time; (ii) interoperability of all entities that contribute to production (customer, services, production systems, etc.); and (iii) extraction of the optimal value-added flow at any time from the data [5].

A smart factory is an environment that consists of Big Data sources, such as data warehouses, Data Lakes (DLs), and databases, whether on-premises or on the cloud, that can produce massive volumes of textual content (unstructured, semi-structured, and structured), multimedia content (images, video, and audio), utilizing a variety of platforms (enterprise, social media, and sensors) during the production cycle. Despite the great and drastic solutions proposed in recent years in the area of Big Data Processing and Systems of Deep Insight, treating Big Data produced by multiple heterogeneous data sources remains a challenging and unsolved problem [4].

Nowadays, in the era of Big Data, with huge amounts of information produced and consumed, data is considered as the “power” of businesses only if is properly processed to offer mainly decision support. Most companies have a lot of unused data that can be used for process mining. This is a side-effect of the widespread digitization and automation of business processes, which leaves digital traces of real process executions as a byproduct. To the best of our knowledge, the integration of DL with process mining activities has not received much attention in research literature yet. As DLs appear to be a promising technique for temporal Big Data storing, the present work, apart from the goals described below, intends to cover this gap as well.

This paper introduces a new approach to handle Big Data in terms of storing and retrieval, which intends to serve best process mining activities that use this data. More specifically, a novel metadata mechanism is proposed that provides the ability to characterize and describe data sources, data items and process related information which are stored in a DL by means of blueprinting. The proposed approach is an extension of prior work on the topic [14], which builds upon the notions of DL and blueprints [16] to add the following contribution: (a) a separate class of blueprints to account for the information related to process mining activities in a smart manufacturing environment (processes, events, machines); (b) the actions to store, retrieve and process data produced by various sources (e.g., sensors) and relate to workflows and mining activities (e.g., events, sequencing, dependencies, etc.); (c) an extension to the DL architecture where we introduce the notion of data puddles to be used for storing smaller portions of data according to some formatting criterion; and (d) the successful application of the framework on a real-world industrial case study, which, on one hand it is quite rare in literature to report a real business customer to study, and on the other hand it yielded some very interesting findings.

The remainder of the paper is structured as follows: Section II discusses related work and the technical background in the areas of Smart Manufacturing, DLs, and Business Process Mining (BPM). Section III presents the DL source description framework and discusses its main components that prepare the relevant data for process mining. This is followed by Section IV that presents an experimental validation of the proposed framework on a real-world case, namely that of poultry meat production. Finally, Section V concludes the paper and highlights future work directions.

II. TECHNICAL BACKGROUND

Nowadays, industries are being transformed with the rise of disrupting technologies and this transformation is called Industry 4.0 or Smart Manufacturing. Industry 4.0 aims to construct an open and smart manufacturing platform for industrial information applications based on a range of disrupting technologies.

As previously mentioned, Smart Manufacturing consists of new scientific trends worldwide, such as IoT, Big Data analytics, cloud computing, Artificial Intelligence (AI), CPS and other new generation information disrupting technologies. All these technologies are related to the recent technological developments where Internet and its supporting technologies serve as a backbone to integrate human and machine agents, materials, products, production lines and processes, within and beyond organizational boundaries, to form a new kind of intelligent, connected, and agile value chain [6].

The vision of smart manufacturing envisages the use of smart technologies, such as information technology, sensor networks, process analysis, and production management and control software, to improve efficiency on agility, asset utilization and sustainability [7].

Industry 4.0 aims at digitizing engineering, production and manufacturing with the goal of:

- A seamless integration device, sensors, machines, as well as software and IT systems
- Increased flexibility thanks to pushing more intelligence from centralized planning systems to the edge
- Increased efficiency due to automated data exchange and analysis within the value chain

The most popular advantages of Smart Manufacturing are [8]:

- Optimization
- Customization
- Cost reduction
- Efficiency
- Customer Experience

The aforementioned information technologies, which are the backbone of smart manufacturing, generate a large volume of heterogeneous data, structured, semi-structured and unstructured. Big data analysis models and algorithms may be executed to organize, analyze and mine this raw data to obtain valuable knowledge. This manufacturing data is collected usually in real-time and sometimes automatically by IoT devices. Manufacturers aim to find a way to increase the efficiency, manage the storage of all this data and visualize it to improve and increase productivity and quality of products.

Most of the work on Big Data integration has been focused on the problem of processing very large amounts of data, extracting information from multiple, possibly conflicting data sources, reconciling the values and providing unified access to data residing in multiple, autonomous data sources.

Various studies addressed isolated aspects of data source management relying on schema mapping and semantic integration of different sources [9][10]. Those studies focused primarily on the construction of a global schema or a knowledge base to describe the domain of the data sources. Most of the proposed techniques examine user queries and return tables related to specific keywords presented in the query; however, keyword-based techniques fail to capture the semantics of natural language, i.e., the intentions of the users, and thus they can only go as far as giving relevant hints.

DL is a storage repository that can store large amounts of structured, semi-structured, and unstructured data. With this Big Data storage architecture, data holders can store every type of data in its native format without the need to structure or clean them until it needed. Every data element in a DL needs to be given a unique identifier and tagged with a set of metadata information [15].

Process mining is an emerging research discipline that helps organizations discover and analyze business processes based on raw event data. Basically, it sits between computational intelligence and data mining on one hand, and process modeling and analysis on the other [12]. Many researchers are developing new and more powerful process mining techniques and software vendors are incorporating these in their software and especially for Big Data [13]. Generally, process mining techniques based on the business log files produced. In our paper we are trying to utilize also Big Data produced by a manufacturing environment as a goal to transform them by the metadata mechanism to participate also in the process mining. This is very important in a world in which data is produced by a vast number of heterogeneous data sources.

There are three types of process mining activities, discovery, conformance checking, and enhancement. These activities use an existing process model produced based on event logs (see Figure 2). Companies and organizations tend to produce their log files according to their own data standards. Therefore, a standardization model is needed, to unify and formalize the description of all business entities in the enterprise under analysis, allowing to efficiently monitor and extract knowledge from event logs. In our case, this standardization is provided through the theory of Blueprint Models.

The target here is to build a unified DL-based business information standardization model, which is tailored to the needs of manufacturing organizations and consists of a

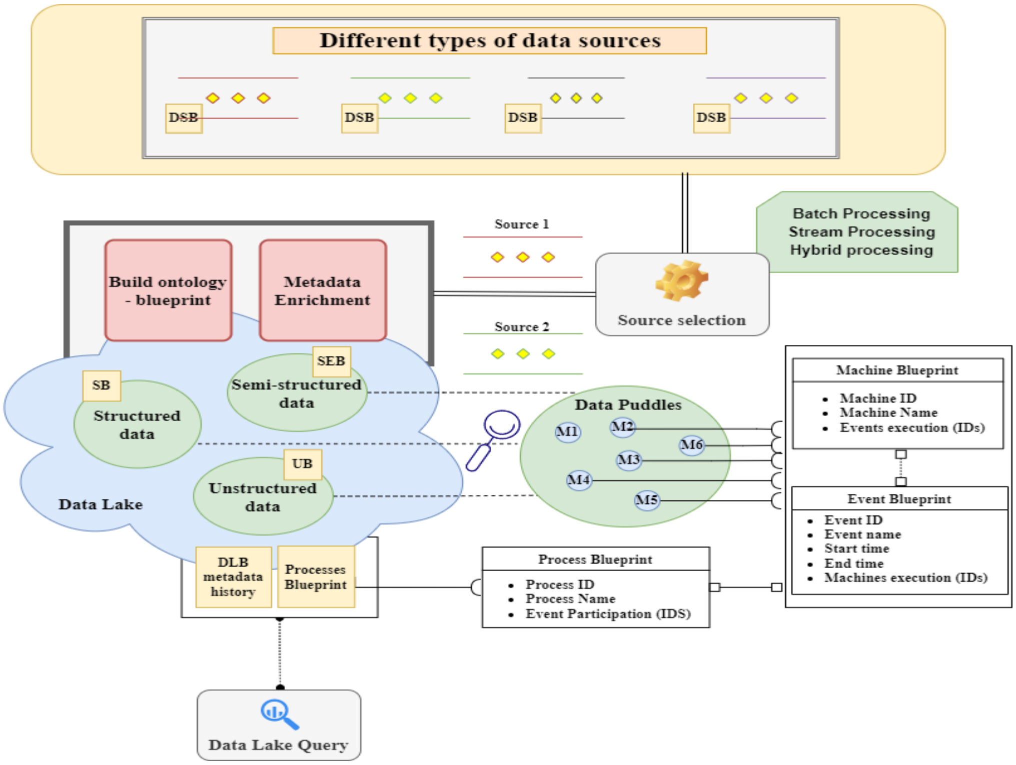


Figure 1. The architectural structure of the proposed approach

number of blueprint entities. These blueprints essentially describe an environment that produces large amounts of different types of data in a specific, disciplined form, including the data sources and their outputs, as well as the business processes. The entities in the business process related blueprints essentially describe and correlate the process information stored in the DL, offering also links to events interacting in chronological order and based on dependencies. These special blueprints codify, integrate, and contextualize business data and processes. They provide parameterized solution-aware patterns that represent operational processes and inter-relate a variety of data of diverse data types, critical functional, sensor, and performance factors in business production. These smart manufacturing intelligence blueprints are built using ontologies and utilizing a dedicated blueprint processing data mechanism along with event logs to facilitate efficient execution of process discovery, conformance checking, and model enhancement.

III. METHODOLOGY

As mentioned above, an extended, unified standardization framework for smart manufacturing and business process related data residing in a DL is introduced in this work, which utilizes a semantic metadata enrichment mechanism via Blueprints [14]. The latter utilizes the 5Vs Big Data characteristics and ontologies to assist data processing (storing and retrieval) in DLs with pond architecture, with emphasis on organizing and preparing data to facilitate process mining.

According to the proposed pond architecture, a DL consists of a set of data ponds, and each pond hosts / refers to a specific data type: structured, semi-structured and unstructured. Each pond contains a specialized storage and data processing system depending on the data type [11].

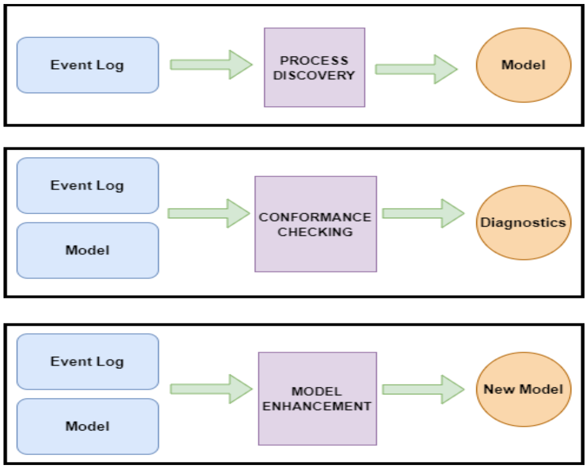


Figure 2. The three types of process mining activities

Process mining is performed mainly by using timestamped data logs. In a DL, however, there are various types of unstructured and semi-structured data, such as images, video, and sounds that may lack time information. Furthermore, structured data as well may not be ready to participate in process mining activities, mainly because it does not have timestamps. To achieve a uniform and constrained approach to the way related data is stored, we will adopt the data blueprint for DL approach created in our previous work (see Figure 1) and extend it by creating three new manufacturing blueprints that describe the data produced by machines and processes in a factory. This is performed by enriching the metadata manufacturing semantics of the DL framework that will prepare the data to be used by process mining tasks.

As mentioned above, this paper builds upon an existing framework [14] which is based on a metadata semantic enrichment mechanism that uses the notion of blueprints [16] to produce and organize meta-information related to each source producing data to be hosted in a DL. In this context, each data source is described via two types of blueprints as shown in Figure 3, which essentially utilize the 5Vs Big Data characteristics Volume, Velocity, Variety, Veracity and Value. The first includes information that is stable over time, such as, the name of the source and its velocity of data production. The second involves descriptors that vary as data is produced by the source in the course of time, such as the volume and date/time of production. The combination of these blueprints creates the Data Source Blueprint (DSB).

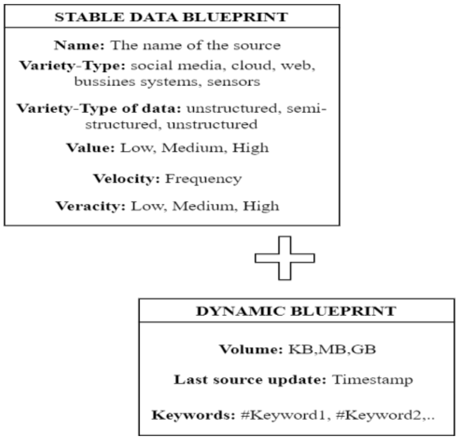

STABLE DATA BLUEPRINT
Name: The name of the source
Variety-Type: social media, cloud, web, bussines systems, sensors
Variety-Type of data: unstructured, semi-structured, unstructured
Value: Low, Medium, High
Velocity: Frequency
Veracity: Low, Medium, High
DYNAMIC BLUEPRINT
Volume: KB,MB,GB
Last source update: Timestamp
Keywords: #Keyword1, #Keyword2,..

Figure 3. Stable and Data Blueprint

As shown in Figure 1, every time data sources or data is pushed in and out of the DL (for example, Source 1 and Source 2, which are accompanied by their DSBs), the stable and dynamic blueprints are updated thus keeping a sort of history of these transactions on the Data Lake Blueprint (DLB) metadata history, which include the Structured Data Blueprint (SB), the Semi-structured Data Blueprint (SEB), and the Unstructured Data Blueprint (UB) residing in the data ponds.

Essentially, the description of the sources helps to treat and manage many, multiple, and different types of data sources and to contribute to the DLs' metadata enrichment before and after these sources become its members. When a data source becomes part of the DL, the metadata schema is utilized, describing the whole DL ontology. The filtering and retrieval of data is based on this metadata mechanism which involves attributes from the 5Vs, such as last source updates and keywords.

The purpose of this paper is to extend the proposed framework via creating in the ponds the so-called data puddles, which are smaller, pre-build datasets as shown in Figure 1, which store data that machines produce in the production line (Machine and Event Blueprints). Furthermore, the existing framework is extended to include a process-related blueprint which provides information about the participation of each machine in various processes during production, that is, which machine executes each event within a process cycle.

To test that the proposed framework works properly and that it meets the needs of a real factory, we investigated its applicability to a major local industrial player, namely Paradisiotis Group (PARG) [17]. PARG is one of the most significant local companies and experts in the field of poultry farming and trading of poultry meat in Cyprus. It offers a wide selection of high-quality products that meet the needs of modern consumers for convenience in cooking and healthy eating. The business processes and the manufacturing data of the factory are, of course, confidential, and therefore, this paper reports only a part of the processes with not so many details, associated with a masked and downgraded version of the data. Nevertheless, the case-study is more than enough to demonstrate the basic principles of the proposed framework and prove its applicability and usefulness.

Every process in the manufacturing cycle consists of events and each event is executed by a machine that participates in a specific blueprint. Figure 4 describes an example of a process followed during the production of chicken nuggets at the PARG factory. First, the ingredients are prepared consisting of 85% chicken breast meat and 15% chicken skin. These ingredients are pushed into a flaker machine that cools down the raw material to a uniform temperature ranging between -2°C and 0°C and then shapes it to blocks. Subsequently, the blocks are chopped in smaller pieces of 8mm size by a mincing machine. Then, these minced pieces are mixed with dry ingredients (such as spices) and water, for 3-5 minutes and the end-product created consists of chicken nuggets formed to have a net weight of 40g each. Finally, the chicken nuggets are deep fried and packaged with appropriate labeling.

The process analyzed above is practically followed for all pre-fried products, such as drumsticks and meatballs, with the only difference being in the forming, with size and shape changing accordingly depending on the product. In addition, for fresh products, such as burgers, the forming event is omitted and another event is added before downcooling, namely the deboning of raw material which is executed by the Tappler machine (out of scope of the present study). In all these processes, the material is pushed from machine to machine via conveyors.

If we analyze the process of producing chicken nuggets depicted in Figure 4, we may derive that the following seven events take place:

- Downcooling *(ID: 12)*
- Chopping 8m *(ID: 4)*

- Mixing *(ID: 7)*
- Forming *(ID: 5)*
- Frying *(ID: 2)*
- Packaging *(ID: 3)*
- Labeling *(ID: 9)*

This process consists of events that are carried-out by machines which produce data during execution. To prepare this data for process mining, we use the process blueprint that provides information about production and is part of the manufacturing DL, as shown in Figure 1, as well as a machine blueprint and an event blueprint, which describe the data that machines produce during the production cycle.

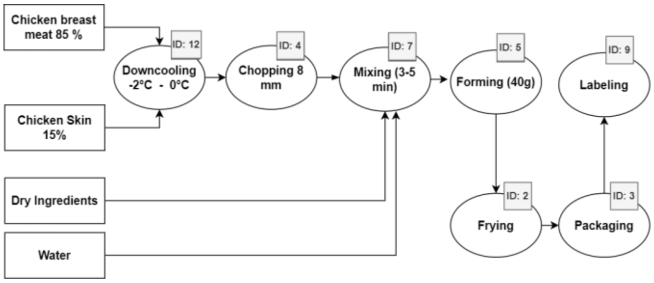


Figure 4. PARG chicken nuggets manufacturing process

As we can see on the previously described process, 10 machines (3 of which perform the same task) participate in the chicken nuggets production:

- Flaker *(ID, Type: FL2, Downcooling)*
- Mincer *(ID, Type: MC1, Chopping)*
- Mixer *(ID, Type: MX3, Mixing)*
- Former (*ID, Type: FRM1, Forming)*
- Fryer (*ID, Type: FR4, Frying)*
- Labeler *(ID, Type: LBI, Labeling)*
- Packager *(ID, Type: PC3, Labeling)*
- Conveyors (x3) *(ID, Type: CV1, CV2, CV3 Conveyor)*

These machines execute multiple events in a specific order during the production of nuggets. The type of machine that executes a specific event is stored in the Manufacturing Blueprint thus being able to check the availability of machines of this type.

The Process Blueprint captures the *Process ID*, the *Process name* and *Events participation*. For example, the chicken nuggets process blueprint is described as follows:

- Process ID: 100
- Process Name: Nuggets production
- Events execution: 12, 4, 7, 5, 2, 3, 9

The Event Blueprint consists of *Event ID*, *Event name*, *Start Time*, *End Time*, *Expected execution time*, *Executed by* (Machine Type) and *Dependencies*, the latter describing what other event has to be executed before the current event may start. For example, the Mixing event is described by the following Event Blueprint:

- Event ID: 7
- Event name: Mixing nuggets ingredients
- Start time: Timestamp
- End time: Timestamp
- Expected execution time: 4 minutes
- Executed by (ID, Type): MX
- Dependencies: 12, 4

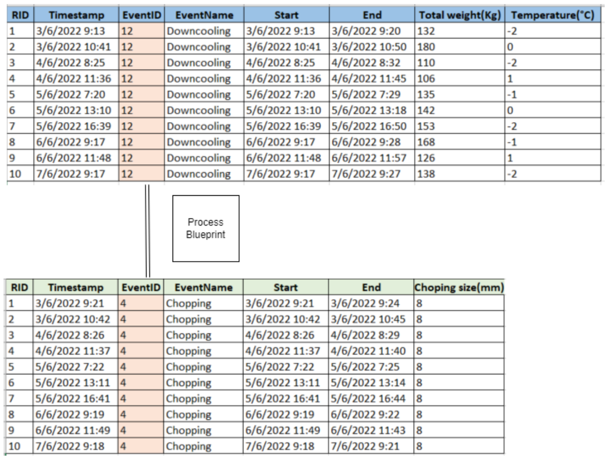
Furthermore, the Event Blueprint captures the *Expected execution time* of the event to be able to check for abnormalities in case the execution of the event is delayed for some reason. This is performed through the analysis of the start and end times, the resources utilized, the roles etc., so as to trace the causes of this delay.

The Machine Blueprint captures *Machine ID*, *Machine Type*, *Machine name* and the *IDs* of the machines that may execute the specific event.

Essentially, the proposed information structure for the description of the data sources that exist in a smart factory efficiently supports the management of multiple data formats. It also allows data to be prepared for process mining through the metadata semantic enrichment that requires events to be timestamped and set in chronological order according to the process executed. Finally, sources that produce unstructured and semi-structured data that are stored in the relevant pond of the proposed approach may also be linked with the rest of the event information and provide added value to the analysis of a certain process. For example, data from a sensor installed on some machine in the production line (e.g., counting packages in the case of chicken nuggets) is coupled with photos captured of a certain spot (e.g., when packaging or labeling) to allow for assessing productivity or the level of defects, offering a complete root-cause analysis.

IV. EXPERIMENTATION

To demonstrate the applicability and effectiveness of the proposed framework, we will use the chicken nuggets production process of the PARG factory described in the previous section. The target here is twofold: First, to demonstrate how the proposed approach was used in practice for the PARG case-study and highlight some interesting findings. Second, to make a short assessment of different DL structures, including the proposed one, according to specific metrics and present the results that show the superiority of this approach.



RID	Timestamp	EventID	EventName	Start	End	Total weight(Kg)	Temperature(°C)
1	3/6/2022 9:13	12	Downcooling	3/6/2022 9:13	3/6/2022 9:20	132	-2
2	3/6/2022 10:41	12	Downcooling	3/6/2022 10:41	3/6/2022 10:50	180	0
3	4/6/2022 8:25	12	Downcooling	4/6/2022 8:25	4/6/2022 8:32	110	-2
4	4/6/2022 11:36	12	Downcooling	4/6/2022 11:36	4/6/2022 11:45	106	1
5	5/6/2022 7:20	12	Downcooling	5/6/2022 7:20	5/6/2022 7:29	135	-1
6	5/6/2022 13:10	12	Downcooling	5/6/2022 13:10	5/6/2022 13:18	142	0
7	5/6/2022 16:39	12	Downcooling	5/6/2022 16:39	5/6/2022 16:50	153	-2
8	6/6/2022 9:17	12	Downcooling	6/6/2022 9:17	6/6/2022 9:28	168	-1
9	6/6/2022 11:48	12	Downcooling	6/6/2022 11:48	6/6/2022 11:57	126	1
10	7/6/2022 9:17	12	Downcooling	7/6/2022 9:17	7/6/2022 9:27	138	-2

RID	Timestamp	EventID	EventName	Start	End	Choping size(mm)
1	3/6/2022 9:21	4	Chopping	3/6/2022 9:21	3/6/2022 9:24	8
2	3/6/2022 10:42	4	Chopping	3/6/2022 10:42	3/6/2022 10:45	8
3	4/6/2022 8:26	4	Chopping	4/6/2022 8:26	4/6/2022 8:29	8
4	4/6/2022 11:37	4	Chopping	4/6/2022 11:37	4/6/2022 11:40	8
5	5/6/2022 7:22	4	Chopping	5/6/2022 7:22	5/6/2022 7:25	8
6	5/6/2022 13:11	4	Chopping	5/6/2022 13:11	5/6/2022 13:14	8
7	5/6/2022 16:41	4	Chopping	5/6/2022 16:41	5/6/2022 16:44	8
8	6/6/2022 9:19	4	Chopping	6/6/2022 9:19	6/6/2022 9:22	8
9	6/6/2022 11:49	4	Chopping	6/6/2022 11:49	6/6/2022 11:43	8
10	7/6/2022 9:18	4	Chopping	7/6/2022 9:18	7/6/2022 9:21	8

Figure 5. Indicative data for PARG's chicken nuggets production process

ISBN: 978-1-61208-997-3

Figure 5 presents an excerpt of the data produced by the Flaken and Mincing machines at PARG factory during the manufacturing process presented in Figure 4. This data is stored in the structured data pond of Figure 1. More specifically, each dataset produced by the two different machines is stored using a different puddle within the data pond. During this process other formats of data is generated as well, such as video and images (omitted due to size limitations) and, since these constitute unstructured and XML-based data (semi-structured), data is stored in the respective pond and distinct puddles, respectively.

As presented in the previous sections, according to the process blueprint describing process with ID100, the execution of events is in sequence 12, 4, 7, 5, 2, 3, and 9. In order to retrieve the data for this specific process, the following SPARQL query should be executed:

```
SELECT ? DLsources
  WHERE {
      ? process ID <has ID> 100  }
```

Running this query on the DL blueprint triggers first the retrieval of information on event execution from the process blueprint. Using this information as the basis, all relevant data for this process is retrieved and mapped depending on the order in which events are executed by machines. As shown in Figure 1, the process blueprint is connected with the event blueprint, which provides the expected execution time of each event by a machine, as well as the type of machine that executes this event, while the machine blueprint describes the events that can be executed by each machine. This information was combined with the data retrieved from the appropriate puddles residing in the specific pond of the DL and were made ready for use in the process mining activities that followed, the latter yielding some interesting results. For confidentiality purposes, we report here two of them very abstractly.

A few delays were encountered in some of the steps, which were revealed during this analysis by comparing the expected with the actual execution time. This led to further investigating these delays through the start and end times of the relevant events. It was noticed that optimization in the way the sequence of the execution of tasks (events) by the machines had ample room for improvement in terms of timing: There were delays in commencing operation from a machine after the previous task was finished. This was partly attributed to roles and resources within the production line, as various tasks were shared amongst employees and, in some cases, multitasking was an issue. All the above were communicated to the senior management of PARG who acknowledged their value for future actions.

The analysis was also supported by unstructured data (images of the production line) which was acknowledged by all parties involved (analysts, managers, workers) to have played a crucial role in identifying bottlenecks. Therefore, as the proposed framework allows for utilizing both unstructured and semi-structured data for process mining, it was considered a significant benefit.

The second experimental aim is to investigate in general the readiness of the manufacturing data residing in a DL to participate in process mining tasks, when the DL has the following structure:

- without the proposed metadata enrichment mechanism
- with the metadata mechanism without a pond architecture
- with metadata mechanism and a pond architecture
- with metadata mechanism, and with pond and puddle architecture (the proposal of this paper)

The following characteristics/metrics are utilized for each DL structure:

- Granularity
- Ease of storing/retrieval
- Process mining readiness
- Expandability

We define *Granularity* [14] as the ability to refine the type of information that needs to be retrieved for a specific factory process. This ability is expressed by the number of fine-grained levels the DL mechanism supports for defining the information sought. *Ease of storing/retrieval* refers to the ability of the DL structure to store or retrieve data in the DL in a simple and easy way. It is assumed here that the retrieval action is efficient enough to return the desired parts of the information sought. This characteristic is reflected on the number of steps that need to be executed for the process of storing and retrieving data items to be completed. Moreover, we define *Expandability* as the ability to expand the structure of the DL and the metadata mechanism with further functional characteristics or other supporting techniques and approaches, such as visual querying and blockchain. Obviously, the more open the mechanism for expansion, the better. Finally, *Process mining readiness* is reflected in the number of steps that need to be executed after the query is executed for the data to be fed to process mining activities. The aforementioned characteristics are evaluated using a Likert linguistic scale, including the values Low, Medium, and High. Table 1 provides a definition of Low, Medium and High for each characteristic introduced.

TABLE 1. DEFINITION OF LOW, MEDIUM, AND HIGH OF EACH ASSESMENT CHARACTERISTIC

Characteristic	**Low**	**Medium**	**High**
Granularity	1 level	2 levels	3 or more levels
Ease of storing /retrieval	5 or more actions	3-4 actions	2 actions maximum
Expandability	No or limited	Normal	Unlimited
Process mining readiness	4 – 5 actions	2 – 3 actions	1 action maximum

As an example, let us now assume that the PARG factory's DL owner wants to retrieve all data present in the DL relevant to the process of chicken nuggets production presented in the previous section so as to feed it to the process mining stages of discovery, conformance checking, and model enhancement. Note that the PARG factory consists of hundreds of production processes. In order to retrieve the data, the following SPARQL query should be formed and executed:

SELECT ? Dlsources
WHERE { process ID <has ID> 20 }

TABLE 2. EVALUATION AND COMPARISON OF DL STRUCTURES

Approach	Granu-larity	Ease of storing /retriev.	Process mining readin.	Expanda-bility
Without metadata mechanism	Low	Low	Low	Low
With metadata mechanism without pond architecture	Medium	Medium	Low	Unlimited
With metadata mechanism with pond architecture	High	High	Medium	Unlimited
With metadata mechanism with pond and puddle architecture	High	High	High	Unlimited

The metadata mechanism with pond architecture (third from top in Table 2) may be considered as the benchmark of our comparison [14]. It presents High *Granularity*, High *Ease of storing/retrieval* using the stable and dynamic data source blueprint descriptions, with a Medium *Process Mining Readiness*, and High *Expandability*. These values are attributed as follows: High *Granularity* is achieved using keywords that essentially describe the sources and the blueprint values. This enables the user to define details at the level of the properties offered, enabling the retrieval of data based on fine-grained query-like information. The High *Ease of storing/retrieval* is achieved by the DL metadata history which stores the blueprint description of the DL as each time data sources are pushed into it. This helps the mechanism to place the sources to a specific pond according to the structure of the data involved (structured, semi-structured and unstructured). This source distribution in the DL also facilitates simple and *Easy Storing and Retrieval* of the information stored. In addition, the implementation of this DL architecture is based on the Hadoop ecosystem and hence this provides High *Expandability*. Finally, the metadata mechanism provides Medium Level of *Process Mining Readiness* due to the lack of defining dependencies and describing the process, events and machines in the production line.

It is logical that, as we move to the upper structural forms of Table 2, the evaluation of the selected characteristics gets worse: Assuming that PARG's DL has an architecture without metadata, a SPARQL query could not be executed at all. In addition, with this structure, all the data is pushed to the DL without any management policy and as a result the DL is highly likely to transform into a Data Swamp, while at the same time it would take quite a few actions (more than five) to retrieve data because all datasets will need to be visited and checked if they are related to the specific process. *Process Mining Readiness* is Low as no clear separation of events, type, dependencies etc., exists, let alone the fact that data needs to be timestamped. Finally, in the absence of a management mechanism, *Expandability* may be characterized as Low. Taking into consideration now that the PARG DL has a structure with the proposed metadata enrichment mechanism but without a pond architecture, the data is pushed to the DL following a metadata policy. As a result, this DL can be characterized with Medium *Granularity* due to the fact that the metadata mechanism provides 2 levels of Granularity, which are provided by the metadata mechanism and the pond the data is stored in, with Low to Medium *Ease of storing/retrieval* as the data is pushed to the DL with its metadata. Therefore, in order to retrieve it, one needs to access and process the metadata to check if a certain piece of information is related to the specific process examined. This DL structure could be characterized also with Low *Process Mining Readiness* because, after executing a query, the data is not separated according to its type and an additional task to separate and timestamp it should be performed, along with proper definition of any dependencies. Finally, it can be characterized with *Unlimited Expandability* as the metadata mechanism allows for practically any extension.

The proposed metadata mechanism with pond and puddle architecture can be characterized similarly to the previous benchmark (3rd row in Table 2), but with High *Process Mining Readiness*. By extending the mechanism reported in [14] and introducing the process blueprint that captures the specific events triggered while a process is executed, the event blueprint that captures the type of machines that participate in a process and the machines blueprint that captures the events that specific machines can trigger, results in a data environment ready to perform process mining readiness. Furthermore, by extending data ponds with data puddles where each puddle stores data from each machine on PARG factory, enables a query to provide the requested data of a specific process to be separated according to its format types.

Table 2 sums up all the information of the short comparison between the DL structures made in this section. It is evident that the proposed mechanism outperforms all alternatives in terms of the *Process mining readiness* characteristic as a result of the extensions made in the blueprints.

V. CONCLUSIONS

This paper proposed a novel smart manufacturing DL framework for process mining utilizing a semantic enrichment mechanism via metadata blueprints [14]. The framework utilizes the 5Vs Big Data characteristics and blueprint ontologies to assist data processing (storing and retrieval) in DLs, the latter being organized with a pond architecture that hosts different types of data, structured, semi-structured and unstructured, enhanced by data puddles. The puddles consist of data produced by machines in the production line and essentially prepare the data in the ponds for process mining activities.

The applicability of the framework was demonstrated and assessed through a real-world case-study on a local poultry meat production factory. The process of producing chicken nuggets was modeled with relevant data captured, stored and processed. Process mining revealed delays and bottlenecks in the sequencing of the execution of events by machines and personnel which may be avoided by optimizing task sharing amongst roles. The senior management of the factory greatly appreciated the support of the proposed approach for decision support with respect to production control.

Furthermore, a short comparison with different DL structures was performed revealing the high potential of the proposed approach as it offers a more complete characterization of the data sources and covers a set of key features reported in literature. Especially, the inclusion of data paddles can greatly enhance the management of manufacturing data that can later participate in process mining activities, such as discovery, conformance checking, and model enhancement utilizing all available data types.

Future research steps will include the full implementation of the proposed mechanism in cooperation with the industrial partner using the metadata model described in this work and extending its application in the context of structured, semi-structured and unstructured data present in the processes of the factory. This will allow the evaluation of the proposed framework in more detail and performing further process mining steps utilizing real-world manufacturing data. As a result, investigation of how to improve privacy, security, and data governance in DLs will be made feasible, also by extending it to include blockchain technology characteristics and smart contracts.

ACKNOWLEDGMENT

This paper is part of the outcomes of the Coordination and Support Actions (CSA) Twinning project DESTINI. This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No.945357.

REFERENCES

[1] F. Tao, Q. Qi, A. Liu, and A. Kusiak, “Data-driven smart manufacturing,” J. Manuf. Syst., vol. 48, pp. 157–169, 2018.J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, pp. 68–73, 1892.

[2] P. Zheng et al., “Smart manufacturing systems for Industry 4.0: Conceptual framework, scenarios, and future perspectives,” Front. Mech. Eng., vol. 13, no. 2, pp. 137–150, 2018. L. Wang, M. Törngren, and M. Onori, “Current status and advancement of cyber-physical systems in manufacturing,” J. Manuf. Syst., vol. 37, pp. 517–527, 2015.

[3] J. Wan et al., “Software-Defined Industrial Internet of Things in the Context of Industry 4.0,” IEEE Sens. J., vol. 16, no. 20, pp. 7373–7380, 2016.D. Kornack and P. Rakic, “Cell Proliferation without Neurogenesis in Adult Primate Neocortex,” Science, vol. 294, pp. 2127-2130, Dec. 2001, doi:10.1126/science.1065467.

[4] M. Farid, A. Roatiş, I. F. Ilyas, H. F. Hoffmann, and X. Chu, “CLAMS: Bringing quality to data lakes,” Proc. ACM SIGMOD Int. Conf. Manag. Data, vol. 26-June-20, pp. 2089–2092, 2016.

[5] S. Erol, A. Jäger, P. Hold, K. Ott, and W. Sihn, “Tangible Industry 4.0: A Scenario-Based Approach to Learning for the Future of Production,” Procedia CIRP, vol. 54, pp. 13–18, 2016.

[6] G. Shao, S. J. Shin, and S. Jain, “Data analytics using simulation for smart manufacturing,” Proc. - Winter Simul. Conf., vol. 2015-Janua, no. Smlc 2012, pp. 2192–2203, 2015.

[7] N. Petersen, M. Galkin, C. Lange, S. Lohmann, and S. Auer, “Monitoring and automating factories using semantic models,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 10055 LNCS, pp. 315–330, 2016.

[8] T. Wuest, D. Weimer, C. Irgens, and K. D. Thoben, “Machine learning in manufacturing: Advantages, challenges, and applications,” Prod. Manuf. Res., vol. 4, no. 1, pp. 23–45, 2016.

[9] M. J. Cafarella, A. Halevy, and N. Khoussainova, “Data integration for the relational web,” Proc. VLDB Endow., vol. 2, no. 1, pp. 1090–1101, 2009.

[10] J. Fan, M. Lu, B. C. Ooi, W. C. Tan, and M. Zhang, “A hybrid machine-crowdsourcing system for matching web tables,” Proc. - Int. Conf. Data Eng., pp. 976–987, 2014.

[11] P. N. Sawadogo, É. Scholly, C. Favre, É. Ferey, S. Loudcher, and J. Darmont, “Metadata Systems for Data Lakes: Models and Features,” in Communications in Computer and Information Science, 2019, vol. 1064, pp. 440–451.

[12] W. M. P. van der Aalst et al., “Business process mining: An industrial application,” Inf. Syst., vol. 32, no. 5, pp. 713–732, 2007.

[13] W. M. P. Van Der Aalst and S. Dustdar, “Process mining put into context,” IEEE Internet Comput., vol. 16, no. 1, pp. 82–86, 2012.

[14] M. Pingos and A. Andreou, “A Data Lake Metadata Enrichment Mechanism via Semantic Blueprints,” In Proceedings of the 17th International Conference on Evaluation of Novel Approaches to Software Engineering - ENASE, ISBN 978-989-758-568-5; ISSN 2184-4895, pages 186-196. DOI: 10.5220/0011080400003176, 2022.

[15] P. Sawadogo and J. Darmont, “On data lake architectures and metadata management,” J. Intell. Inf. Syst., vol. 56, no. 1, pp. 97–120, 2021.

[16] M. P. Papazoglou and A. Elgammal, “The manufacturing blueprint environment: Bringing intelligence into manufacturing,” 2017 Int. Conf. Eng. Technol. Innov. Eng. Technol. Innov. Manag. Beyond 2020 New Challenges, New Approaches, ICE/ITMC 2017 - Proc., vol. 2018-Janua, pp. 750–759, 2018.

[17] “Home - Paradisiotis Group,” Paradisiotis Group (PARG). [Online]. Available: https://paradisiotis.com/. [Accessed: 13-Jul-2022].