Cyprus University of Technology

Software Engineering and Intelligent Information Systems Research Lab

A SMART MANUFACTURING DATA LAKE METADATA FRAMEWORK FOR PROCESS MINING

Michalis Pingos & Andreas S. Andreou

Presenter: Michalis Pingos {michalis.pingos@cut.ac.cy}
Department of Computer Engineering and Informatics,
Cyprus University of Technology

The Seventeenth International Conference on Software Engineering Advances
ICSEA 2022
October 16, 2022 to October 20, 2022 - Lisbon, Portugal

IARIA

SHORT CV

Date of Birth: 09/07/1993

BSc: Electrical Engineering, Computer Engineering and Informatics - CUT

MSc: Data Science and Engineering – CUT

PhD Candidate: Department of Electrical Engineering, Computer Engineering and Informatics - CUT

Working Experience: Software Developer, Associate Researcher, Teaching Assistant, Special Scientist, Key expert at Cyprus Youth Board – Makerspace Larnaka, Computer Engineer and Informatics Officer at Tax Department of the Cyprus Ministry of Finance

OUTLINE

- Introduction
- Technical background
- Related work
- Methodology
- Experimentation
- Conclusion & Future work

INTRODUCTION (1/2)

Research problem:

- Handling multiple heterogeneous data sources of a smart factory
- Massive amounts of structured, semi-structured and unstructured data → need to manage them through Data Lake (DL)
- A metadata semantic enrichment mechanism that enables fast storing and efficient retrieval from a smart factory DL
- Providing the ability to serve best process mining activities

INTRODUCTION (2/2)

Main research contributions :

- Smart Manufacturing Data Lake Metadata Framework for Process Mining
- An extended semantic enrichment standardization framework for storing data/data sources in a smart factory using data blueprints
- Developing an extension to the DL architecture where we introduced the notion of data puddles
- To be used for storing smaller portions of data according to some formatting criterion
- Enhances big data storing and retrieval issues by providing process mining readiness

TECHNICAL BACKGROUND (1/8)

Big Data

- Refers to the large amounts of digital data generated by tools and machines, and the global population
- 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

TECHNICAL BACKGROUND (2/8)

Industry 4.0

- Based on a number of new and innovative technological developments, such as [1]:
  - ❖ Cyber Physical Systems (CPSs)
  - ❖ Internet of Things (IoT)
  - ❖ Cloud Computing, Cognitive Computing
  - ❖ Robotics
  - ❖ Augmented Reality (AR) technology and intelligent tools
- 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 a set of CPSs that will interact with each other

TECHNICAL BACKGROUND (3/8)

Industry 4.0 /CPSs

- A CPS consists of mechanisms controlled or monitored by computer algorithms, integrated into the Internet and its users
- Change the way people interact with machines
- 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]
- A smart factory is an environment that consists of Big Data sources

TECHNICAL BACKGROUND (4/8)

Social Media

Cloud

Web

ID	Name	Surname
12A351	John	Smith
871TY5	Andrew	Mephewy

Structured Data


```
{
 "firstName": "John",
 "lastName": "Smith",
 "isAlive": true,
 "age": 27,
 "address": {
  "streetAddress": "21 2nd Street",
  "city": "New York",
  "state": "NY",
  "postalCode": "10021-3100"
 },
 "phoneNumbers": [
  {
   "type": "home",
   "number": "212 555-1234"
  },
  {
   "type": "office",
   "number": "646 555-4567"
  }
 ],
 "children": [],
 "spouse": null
}
```

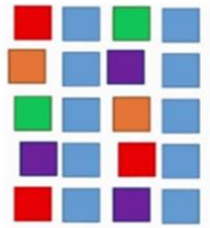

Semi - structured

John Smith is 27 years old. He lives in New York at 21 2nd Street 10021-3100. His phone number is 212 555-1234 and his office number is 646 555-4567. He doesn't have any children.

Unstructured Data

Business Systems

Sensors



Different types of data sources and data formats

TECHNICAL BACKGROUND (5/8)

Data Lakes (DL)

- Big Data processing includes the management of multiple and various types of data: structured, semi-structured, and unstructured
- A storage repository that could store a vast amount of raw data of these various types in its native format and selected and organised when needed
- It is a place to store every type of data in its native format with no fixed limits on account size or file
- Offers high data quantity to increase analytic performance and native integration

TECHNICAL BACKGROUND (6/8)

Data Lakes (DL) (… continued)

- A quite new data storage architecture linked with Big Data processing with unsolved challenging problems [4]
- Used to store large amounts of relational and non-relational data
- Two of the major and challenging problems of DL:

(i) no descriptive metadata or mechanism to maintain metadata leading to data swamp [5]

(ii) security (privacy and regulatory requirements) and access control as data in a folders

TECHNICAL BACKGROUND (7/8)

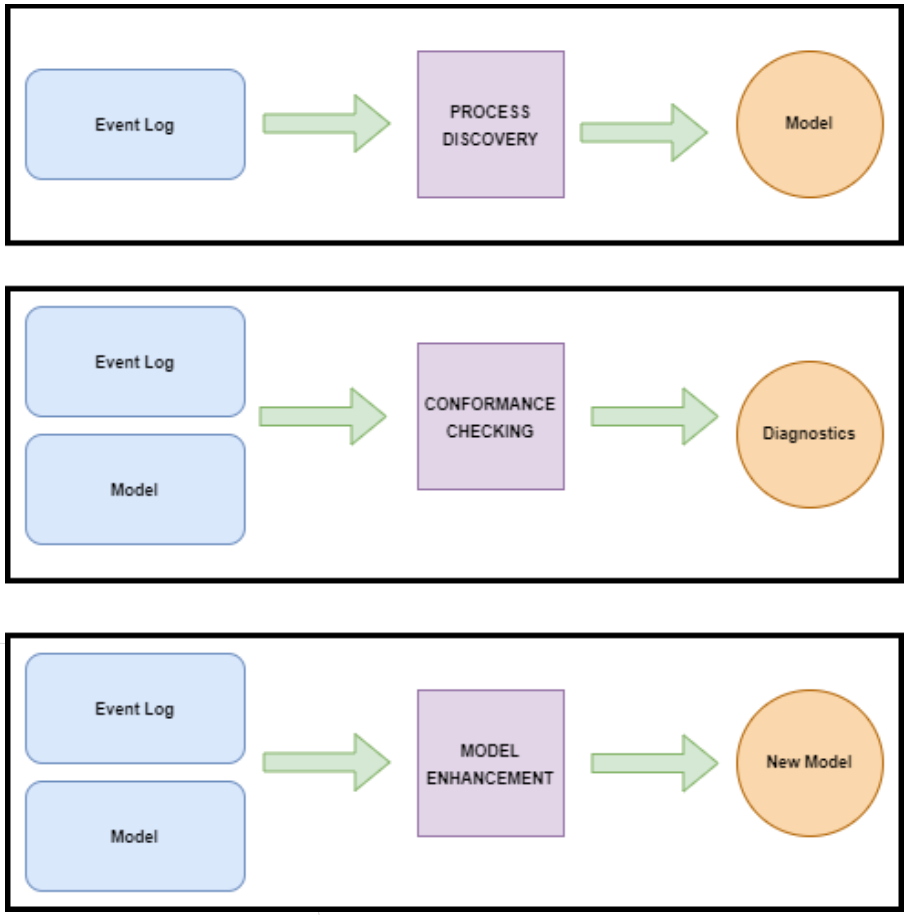
Process Mining

- An emerging research discipline that helps organizations discover and analyze business processes based on raw event data
- 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 now to the world of Big Data [13]
- Generally, process mining techniques based on the business log files produced

TECHNICAL BACKGROUND (8/8)

Process Mining (… continued)

- Companies and organizations tend to produce their log files according to their own data standards
- 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
- In our paper we are trying to utilize also Big Data as a goal to transform them by this standardization
- A big challenge in a world in which data is produced by a vast number of heterogeneous data sources



The three types of process mining activities

RELATED WORK (1/4)

- Sawadogo et al. 2019 [11], identified and presented six main functional characteristics that should ideally be provided by a DL metadata system:
  - Semantic Enrichment (SE)
  - Data Indexing (DI)
  - Link generation and conservation (LG)
  - Data Polymorphism (DP)
  - Data Versioning (DV)
  - Usage Tracking (UT)

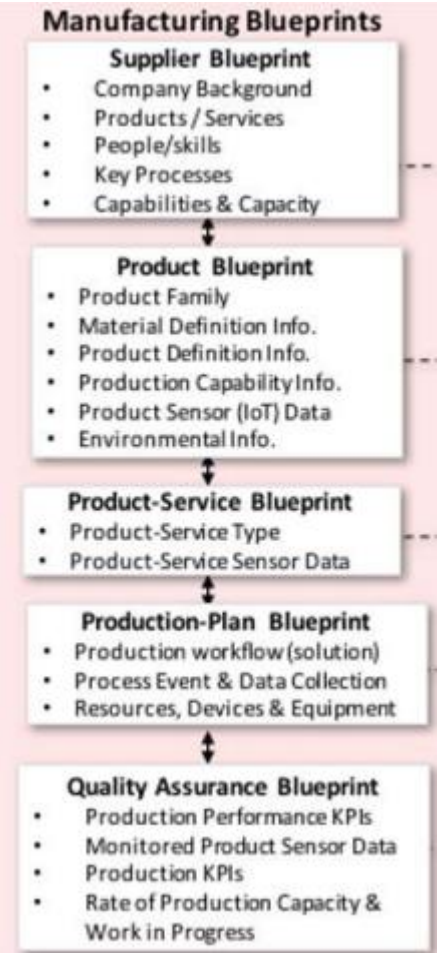
RELATED WORK (2/4)

- Our previous work [14] extended the aforementioned list of characteristics by comparing the metadata mechanism with the two most completed systems: CoreKG and MEDAL [11]. The new list of the characteristics include:
  - ✦ Granularity
  - ✦ Size and type of metadata
  - ✦ Ease of storing/retrieval
  - ✦ Expandability
- None of the existing mentions process mining readiness as a characteristic that can add value to the synthetic examination of the quality and efficiency of metadata enrichment mechanisms for DL

RELATED WORK (3/4)

Manufacturing Blueprints

- Provide a complete summary of a product
- Juxtapose its features with operational and performance characteristics
- Focus → How it is manufactured, which processes are used, and which manufacturing assets are used to make it
- Description → How manufacturers and suppliers coordinate, arrange manufacturing processes, expedite hand-offs, and create the final product [16]

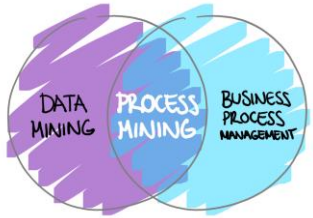


RELATED WORK (4/4)

- This paper extends/enhances our previous work which adopts the basic principles of manufacturing blueprints [14] to provide process mining readiness
- Modifies their purpose and meaning to reflect the description and characterization of sources and the data they produce via the utilization of the five Big Data characteristics
- Describe data sources by means of specific types of blueprints through an ontology-based description representation
- Big Data sources accompanied by a blueprint metadata description before they become part of a DL

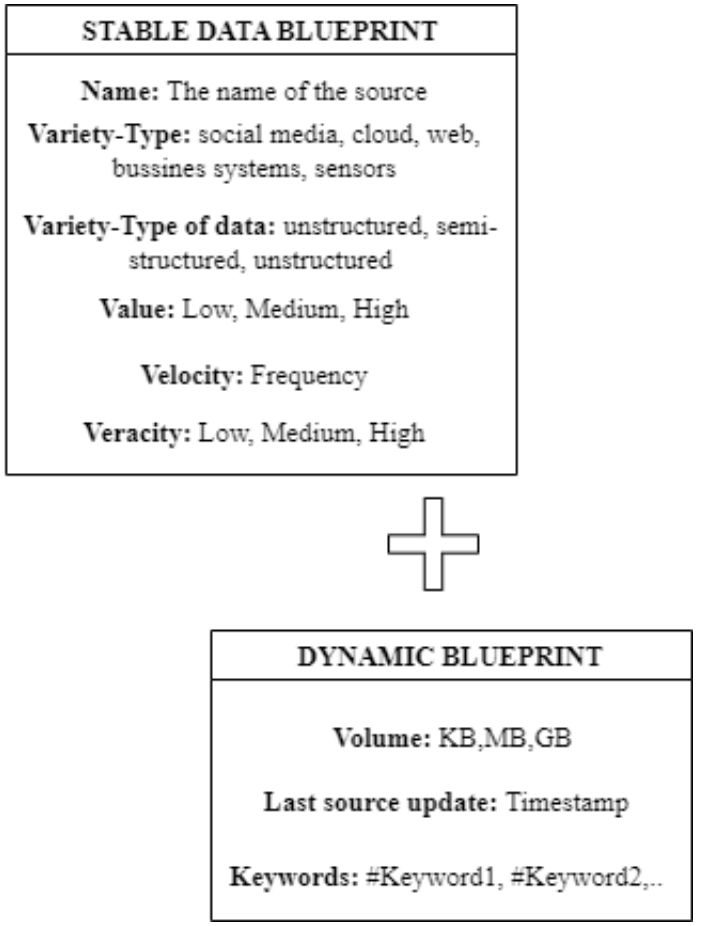
METHODOLOGY (1/7)

- An extended, unified standardization framework for smart manufacturing and business process related data residing
- Utilizes a semantic metadata enrichment mechanism via Blueprints and 5Vs to assist data processing in DLs with pond
- Each pond hosts / refers to a specific data type and contains specialized storage
- Process mining is performed mainly by using timestamped data logs
- In a DL there are various types of data that may lack time information
- May structured data are not ready because not have timestamps



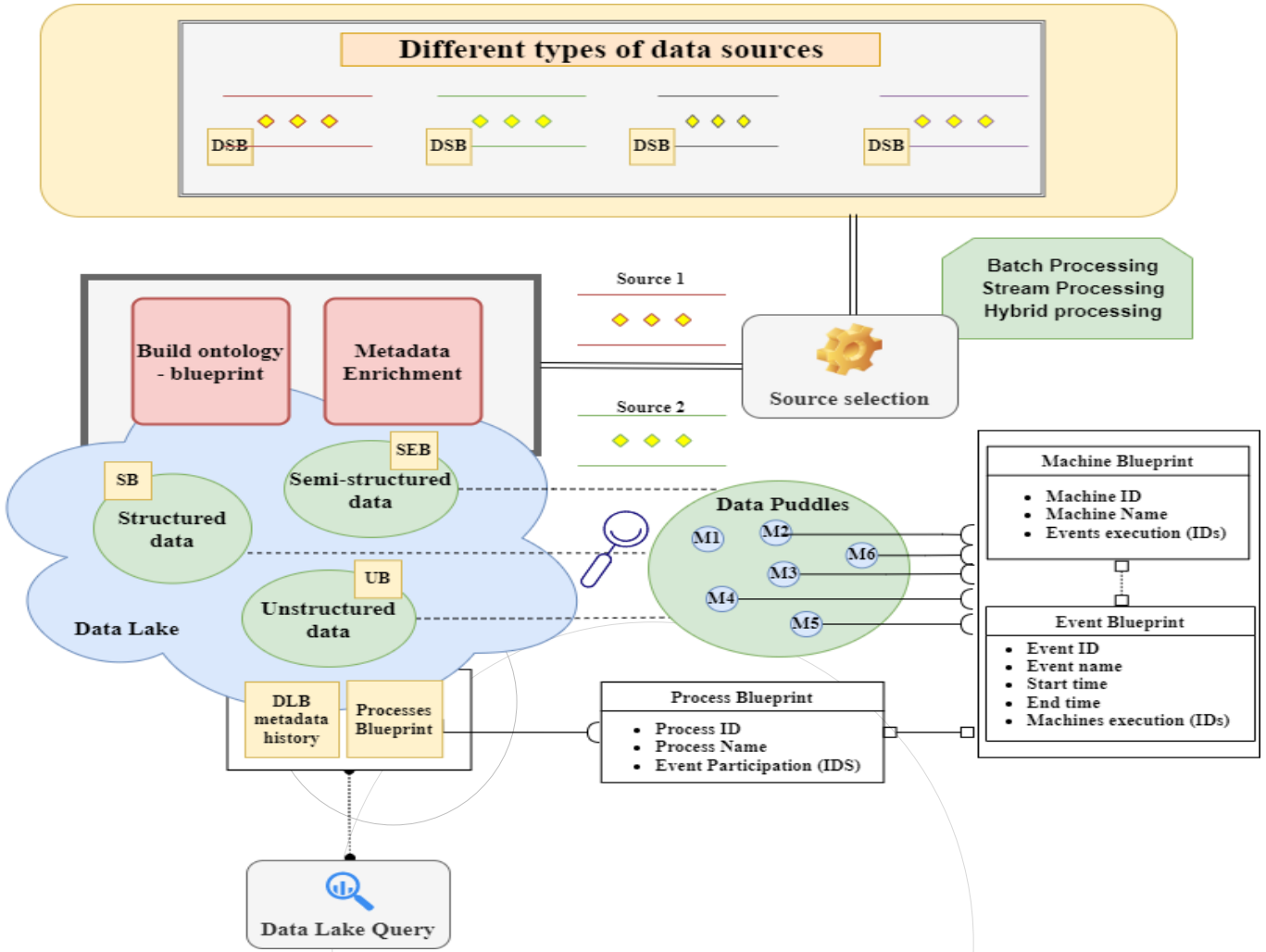
METHODOLOGY (2/7)

- 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
- Each data source is described via two types of blueprints which utilize the 5Vs Big Data characteristics
- The first includes information that is stable over time
- The second involves descriptors that vary as data is produced by the source in the course of time
- The combination of these blueprints creates the Data Source Blueprint (DSB)



Stable and Dynamic Blueprint

METHODOLOGY (3/7)



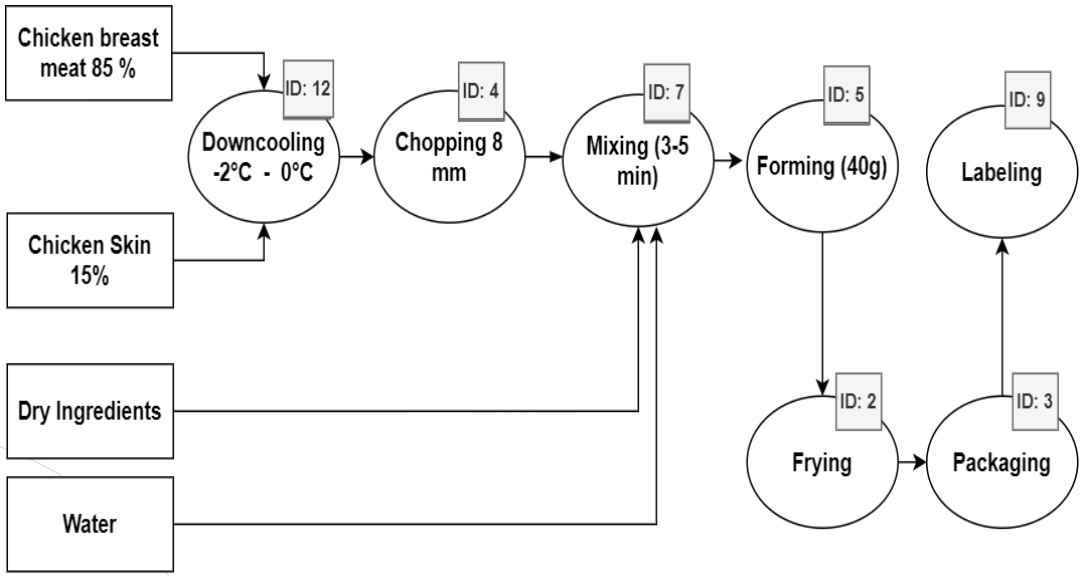
The architectural structure of the proposed approach

METHODOLOGY (4/7)

- Extend the proposed framework via creating data puddles
- Which are smaller, pre-build datasets, which store data that machines produce in the production line (Machine and Event Blueprints)
- Extended to include a process-related blueprint
- Which provides information about the participation of each machine in various processes during production and which machine executes each event within a process cycle
- To test that works properly and meets the needs of a real factory investigated its applicability to a major local industrial player, namely Paradisiotis Group (PARG)

METHODOLOGY (5/7)

- Every process in the manufacturing cycle consists of events and each event is executed by a machine that participates in a specific blueprint
- An example of a process followed during the production of chicken nuggets at the PARG factory
- The process analyzed is practically followed for all pre-fried products
- Every process in the manufacturing cycle consists of events and each event is executed by a machine that participates in a specific blueprint.



PARG chicken nuggets manufacturing process

METHODOLOGY (6/7)

Process

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

Events that take place

- Downcooling *(ID: 12)*
- Chopping 8m *(ID: 4)*
- Mixing *(ID: 7)*
- Forming *(ID: 5)*
- Frying *(ID: 2)*
- Packaging *(ID: 3)*
- Labeling *(ID: 9)*

Event description

- 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

Machines participate

- 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)*

METHODOLOGY (7/7)

- 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
- 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
- Sources that produce unstructured and semi-structured data that are stored in the relevant pond of the proposed approach linked with the rest of the event information
- Provide added value to the analysis of a certain process

EXPERIMENTATION (1/6)

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

EXPERIMENTATION (2/6)

- Data produced by the Flaken and Mincing machines at PARG factory during the manufacturing process presented
- This data is stored in the structured data pond
- Each dataset produced by the two different machines is stored using a different puddle within the data pond
- Other formats of data is generated as well, such as video and images
- These constitute unstructured and XML-based data (semi-structured)
- Data is stored in the respective pond and distinct puddles according the machine

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

Process Blueprint

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

Indicative data for PARG's chicken nuggets production process

EXPERIMENTATION (3/6)

- In order to retrieve the data for the chicken nuggets process, the following SPARQL query should be executed:

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

- DL blueprint triggers first the retrieval of information on event execution from the process blueprint
- All relevant data for this process is retrieved and mapped depending on the order in which events are executed by machines.
- Process blueprint is connected with the event blueprint and event blueprint with the machine blueprint
- This information was combined with the data retrieved from the appropriate puddles

EXPERIMENTATION (4/6)

- The latter yielding some interesting results
- 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
- Optimization in the way the sequence of the execution of tasks (events) by the machines had ample room for improvement in terms of timing
- 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 the process mining readiness of the manufacturing data residing comparing different structures

PRELIMINARY VALIDATION (5/6)

- *Granularity:* ability to refine the type of information that needs to be retrieved expressed by the number of fine-grained levels the metadata mechanism supports for defining the information sought

- *Ease of storing/retrieval:* ability of the metadata mechanism to store or retrieve data in the DL in a simple and easy way reflected on the number of steps that need to be executed

- *Expandability:* ability to expand the metadata mechanism with further functional characteristics, or the support for inclusion of supporting techniques or approaches, such as visual querying

- *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

PRELIMINARY VALIDATION (6/6)

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

Definition of Low, Medium, and High of each assessment characteristic

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

Evaluation and comparison of DL structures

CONCLUSIONS AND FUTURE WORK (1/3)

- A novel smart manufacturing DL framework for process mining utilizing a semantic enrichment mechanism via metadata blueprints
- 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 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 PARG factory

CONCLUSIONS AND FUTURE WORK (2/3)

- Process mining revealed delays and bottlenecks in the sequencing of the execution of events by machines
- The senior management of the factory greatly appreciated the support of the proposed approach for decision support with respect to production control
- A short comparison with different DL structures was performed revealing the high potential of the proposed approach
- Data paddles can greatly enhance the management of manufacturing data that can later participate in process mining activities utilizing all available data types

CONCLUSIONS AND FUTURE WORK (3/3)

Future work

- Full implementation of the proposed mechanism in cooperation with the industrial partner using the metadata model described
- Extending its application in the context of structured, semi-structured and unstructured data present in the processes of the factory
- Evaluation of the proposed framework in more detail and performing further process mining steps utilizing real-world manufacturing data.
- Investigation of how to improve privacy, security, and data governance using blockchain technology characteristics and smart contracts

IARIA

Cyprus University of Technology

SEiIS
Software Engineering and Intelligent Information Systems Research Lab

THANK YOU

michalis.pingos@cut.ac.cy

http://seiis.cut.ac.cy/pingosseiis.html

This paper is part of the outcomes of the CSA Twinning project DESTINI. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 945357.

SELECTED 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.
- [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.
- [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.