Proceedings of the ASME 2019
International Mechanical Engineering Congress and Exposition
IMECE2019
November 11-14, 2019, Salt Lake City, UT, USA

IMECE2019-11362

A DIGITAL TWIN FOR INTRA-LOGISTICS PROCESS PLANNING FOR THE AUTOMOTIVE SECTOR SUPPORTED BY BIG DATA ANALYTICS

Guilherme Guerreiro, Paulo Figueiras, Ruben Costa[1], Maria Marques
CTS, UNINOVA,
Caparica, Portugal

Diogo Graça, Gisela Garcia
Volkswagen Autoeuropa
Palmela, Portugal

Ricardo Jardim-Gonçalves
Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia
Caparica, Portugal

ABSTRACT

One of the areas that can heavily benefit with Industry 4.0 is the logistics, namely with the association of sensing technologies and the application of techniques such as Big Data Analytic, Data Visualization, prediction algorithms, and especially 3D simulation. The association of real data, prediction techniques, and 3D models, allow the creation of realistic Digital Twins that emulate factory processes, making possible the experimentation and testing of new ideas and different scenarios by tweaking key variables, without stopping production.

However, there are many challenges in order to handle and compute all fast-growing, multi dimension data generated, so that all this production related data can be quickly used for defect control, preventive maintenance, advanced analytics for production and resources management, or even later simulation. The work presented in this paper focus in this "in between" processing work, presenting an easily deployable and self-reconfigurable Big Data architecture, where different technologies can work together to extract, transform, load, apply analytics, and then feed a 3D Digital Simulation model. The work presented in this paper is funded by the EU project BOOST4.0 and focus in a specific logistic process of car manufacturing.

Keywords: Digital Twin, Big Data, Industry 4.0, Swam Architecture, Distributed Processing, Manufacturing Data.

1. INTRODUCTION

Digitalization is arriving at every industry and is expected to become standard in this new era of production manufacturing. Especially pushed by the need to improve production methods and keep companies' competitiveness. Industry 3.0 had the main objective of addressing three main market demands [1]: (i) production volume, (ii) product variety, and (iii) delivery time, with the last one being the real challenge at the time, since Industry 1.0 and Industry 2.0 dealt with the first two (i and ii), though mass production and the lean manufacturing concept. Yet, and much influenced by the reduced life cycles of electronic products, there was a need to have dramatically shorter production windows and increase responsiveness. Industry 3.0 addressed this problem with the introduction of electronics and ICT systems with better reconfigurability. The three Industrial revolutions, and the fourth that is occurring now, can be seen as the application of the technological advancements at that specific point in time [2], that is, Steam machines in Industry 1.0, electricity in Industry 2.0, Information technology and automation in Industry 3.0, and Cyber Physical Systems (CPS) in Industry 4.0. Industry 4.0, pushed by a German initiative in 2012 [3] towards the implementation of CPS [4] in industrial environments, refers to systems that have both computational and physical capabilities, being able to exchange information with other systems in real-time, and ideally, add intelligence and efficiency to industrial processes. CPSs can help automate decision making in the production life cycle of a product, triggering control actions and reporting information.

Today, technology already delivers much of the solutions needed to accomplish Industry 4.0, namely sensing devices, communication technologies, autonomous machines, intelligent actuators, Cloud Computing, between others. Nevertheless, most of the times, the integration of different devices is still missing, while from a systems' perspective, the interoperability

[1] Contact author: rddc@uninova.pt

and integration of CPSs are key to rise the quality of Industry 4.0 [2] applications. This point embodies one of the current challenges, how to properly integrate technologies and take advantage of the acquired data from the different machines and processes, so that is possible to handle data with high volume, high speed, different value, multi-structured, and then apply analytics, learning algorithms to detect patterns, do prediction, and apply advanced visualization techniques, at due time. Moreover, how can all enabling technologies to be deployed together, with easy re-configurable, to fit data processing needs and help the creation of reliable Digital Twins? And since a Digital Twin is a virtual replica of the physical world, the Twin will only be as good as the quality and volume of the data, tough, handling this data is a task especially challenging when dealing with the data characteristics in question [5] [6]. By transforming all this raw data into valuables digital representations, either in charts or 3D simulations, management and planning can be much more knowledgeable and efficient, increasing the capability to respond to different needs and, for instance, to plan quickly reaction to any issue in production by rescheduling tasks, if needed. On the other hand, it also makes possible the creation of future scenarios, considering changes on key factors[7].

This data-driven approach should not be limited to manufacturers' premises and can also integrate automatic data-exchange between the manufacturing company and other parties on which production depends, such as logistic and supply chain entities, in order to improve supply chain efficient [8].

1.1 Context

The study presented in this paper is framed within an European Commission-funded project, called BOOST4.0 – Big Data for Factories [9], whose main goals are to push forward Industry 4.0 implementation across the industry and contribute to novel state of the art data-driven solutions, taking into account real use-case scenarios in different factories placed in several countries.

BOOST4.0 also wants to connect factories through the so-called "industrial internet", tackling the lack of standards in this subject. For this matter, the project uses the European Industrial Data Spaces (EIDS) [10] initiative, which is an open platform to facilitate protected exchange of data between a provider and a user. The platform wants to connect "data related service providers" and factories, having a complete ecosystem under development for that end. Other objectives relate to assuring the needed infrastructures, trusted middleware and certification for equipment. In Figure 1, can be seen the different software components of EIDS and the different roles when, for example, company A wants to exchange information with company B. Being involved internal and external connectors IDS (Industrial Data Space), an AppStore and a Broker.

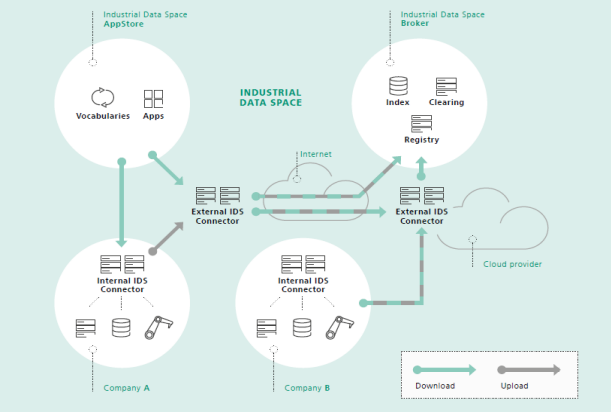

Figure 1 Example of Industrial Data Space Platform data-exchange (Otto et al. 2018).

More explicitly, the work presented here describes an architecture to handle Big Data in all processing stages, from ETL (Extract, Transform, Load) to Data Analytics and visualization. After the data from specific logistic processes is processed and stored, a controller is able to correlate and group the different data sources and coordinate with a 3D Simulation model, using a broker to send data and control the data streams into it. This 3D Digital Twin, on the other hand, besides providing visual meaning for the data, delivers KPIs (Key performance indicators) over specific points of interest, which help assess operation´s efficiency. This 3D model, built in Visual Components (VC) software[11], emulates the scenario, which focus in the batteries intra-logistics processes ,better described ahead in this document. Scenario taken from Volkswagen Autoeuropa pilot, located in Palmela, Portugal. The architecture adopts a distributed and easily scalable philosophy that takes advantage of the Swarm concept applied by Docker [12], for distributed and parallel processing, helping in the deployment of several Big Data technologies, as for instance Apache Spark[14], able to split processing jobs into smaller ones and distribute them in the Swarm Spark workers nodes. Docker Swarm is also key to easily scale up and down resources and technologies without almost any effort for reconfigurability. The architecture also follows the orientation of BOOST4.0, thus, there are APIs ready for a later integration with the EIDS ecosystem. The validation of the approach and Digital Twin is addressed in the specific BOOST4.0 pilot case, with its specific challenges and requirements, and aims to provide all data flow with the necessary technologies to accomplish data related tasks and ingestion of the data into the 3d model, created to access and tests logistic scenarios in processes like sequencing and resources operation. Additionally, the goal is to combine machine learning and forecasting with Simulation, which can help dramatically logistics planning and validation.

This document is organized as follows: related work is presented in section 2. Section 3 presents the methodology, use case scenario and business and technological requirements. In

section 4, it is presented the technical architecture and technology stack used, in section 5 is disclosed how the Digital Twin works, and finally section 6 settles the paper and points the ideas for the future work.

2. RELATED WORK

Studies acknowledge the importance of handling data acquired from Industry 4.0 enabling technologies as raw material and challenge of this evolutionary stage [8] [13]. Another challenge is about the integration of the different CPS with systems in unified architectures, authors in [14], tackle this problems by proposing an unified architecture of 5 levels: Connection, Conversion, Cyber level, Cognition level and Configuration level in order to the flow of data in a CPS enabled factory.

Big Data processing is indicated as the supportive technology to gather, visualize, and then improve production efficiency. As in [15], where it is used to perform quality control on continuous production for fault identification and predictive maintenance. In this matter, other studies indicate Cloud- based approaches with internet services for analytics[16], using, as the authors of [17], similar data pipelines for different use cases, depending on Big Data technologies that enable in-memory, distributed processing. Showing the performance benefits of such approach, besides different deployment configurations. Technologies utilized are Apache Spark, Hadoop, Kafka, Hive, between others. Approach that seems to be equal when dealing with Big Data ,regardless the domain[6].

Digital Twins enrichment with Big Data in smart manufacturing is also a common subject in the literature[7][18], stated as beneficial for more precise, reasonable and more intelligent Twins, being also complementary of each other. Helping in the convergence between the physical and the virtual world.

3. METHODOLOGY AND SCENARIO: BATTERIES PROCESS

CRISP-DM (Cross Industry Standard Process for Data Mining) [19] guidelines were adopted to guide the work and assess pilot´s requirements and needs, either from a business perspective or a technical perspective, and will be followed during pilot´s testing to continuously improve the architecture, if needed. CRISP-DM describes common approaches for data mining related projects, being a well-recognized methodology in the industry. It can be described in the followed steps: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. It is a continuous, not static, process, not ending with deployment.

In the following section, 3.1 and 3.2, the business understanding and data understanding of the scenario are briefly described.

3.1 Scenario and Pilot Use Case

The scenario used is taken from Volkswagen Autoeuropa plant, plant located in near Lisbon, Portugal. Volkswagen Autoeuropa is a high production automotive industry plant, part of Volkswagen Group, with high strategic importance for the country, having 3600 workers alone and being responsible for around 8000 indirect jobs. The plant also works with hundreds of suppliers of all kinds of parts and components, being responsible for a significant exportation volume of the country. The plant is one of the most modern automotive production facilities in Europe, being commonly used to incorporate novel technologies in automation and computerized control.

Regarding the specific scenario itself, it focuses in the generic logistic process flow for the batteries, which are an essential part when assembling cars, and it involves the following processes:

1. Receiving and unloading batteries to a buffer area;
2. Warehousing;
3. Transport the batteries to a sequencing area;
4. Sequence the batteries according to assembly order;
5. Transport parts to the assembly line site (Point of Fit);
6. Assembly the parts.

Some of the processes have systems that help optimize the job, as happen in storage and warehousing, yet, others still depend in a great extent on human intervention, such as to operate some transportation and in the sequencing stage. The main motivations of Volkswagen Autoeuropa pilot is to see how other processes can be also automated, such as using new and smarter AGVs(Autonomous Guided vehicles), better benefit from data generated by systems and properly process and manage and extract information from data, apply new artificial intelligence algorithms, and test realistic 3D simulation to better manage and plan necessities in the assembly line, working towards better adaptability and predictability. The work presented here focus especially on the data-driven architecture and in the 3D Digital Twin that uses the real layout and real data from the factory, to better copycat the reality of processes, such as in the time dimension.

3.2 Data and Technical Requirements

The architecture presented here aims to address all identified technical requirements for the scenario, such as easy deployment, ability to scale up and down attending to processing and storage necessities, have easy reconfigurability to fit production changes, and be able to handle the challenges that data from production systems at site. Being essential the integration of technologies able to collect data from different types of databases, systems, files, streams, etc.

An internal analysis, about data sources and data characteristics was made, concerning the kind of data is being collected, and in what variety and veracity. Characteristics normally assess to identify Big Data [5]. The description of the data is only qualitative and superficial due to data privacy issues.

Resulting in the following:

a. Manufacturing Data: in batch, has medium variety and high veracity;

b. Engineering Data: in batch, has medium variety and high veracity:
c. Product & Process Data: stream data, with high variety and medium veracity;
d. Automated Storage, retrieval of parts: stream data, has high variety and medium veracity;
e. Supply Chain Data: stream data, has high variety with medium veracity;
f. Other Related Data: mainly in batch, presenting high variety and medium veracity.

The volume of datasets is in the order of Terabytes. Besides these basic requirements, the architecture must also be able to accommodate EIDS in the future, and so, have services to get and make data available to future EIDS connectors, providing data interoperability with this platform.

4. ARCHITECTURE

4.1 Description: Layers and Technologies

Figure 2 shows, in its left side, the reference architecture that comprise technologies responsible for all data related processing work, since data collection of data from sensors and CPSs working on premise, to tasks like data harmonization, storage, Data Mining and Machine Learning. In the right side of Figure 2 is a diagram showing how the architecture feeds the 3D Digital Twin, where data is sent into the simulation, from Visual Components, and the feedback (KPIs of processes) received from it, back to system's architecture. The architecture and simulation communicate via FIWARE Orion [20], which is a broker for context information management.

The Data Collection and Ingestion layer is responsible for collecting, extracting, perform data transformation and cleaning to data from the different systems or even from third-party data providers. Data is available normally either from services, files or RDBMS databases. This phase needs to be performed in the most efficient and quick way as possible, in order to deliver near real-time analytics and feed other Intelligent systems properly and at due time, namely when any production rescheduling is needed.

Afterwards, data can be stored using the technology that most fits the end purpose (long-time storage, short term storage for real-time analytics, large batch processing jobs for Data Mining) and characteristics (datatype, schema). NoSQL (Not only SQL) approaches can improve latency when working with multi-structured data, reducing data transformation time.

The Processing layer is where demanding computing tasks are performed, supporting data preparation tasks, transformations, aggregation, mining, between others. These processing work supports the visualization and querying layer, independently of its final form (charts, graphs, simulation). Technologies in this layer are particularly powerful, being able to split big processes and compute them in smaller processing jobs, such as using MapReduce paradigm or the Directed Acyclic Graph concept, to distribute the work across several nodes, so that the processing tasks are made in parallel jobs.

Finally, in the Querying-Analytics-Visualization layer is where are applied Visualization, Analytics, Knowledge Discovery, and Machine learning techniques. Therefore, this layer is responsible for helping transform data into useful and meaningful information in the most easy and utilitarian way. Information that can help improve efficiency along the production supply chain and help decision making, providing findings that can lead to revisions of production plans.

Table 1. Technology stack.

Layer	Technologies
Data Collection and Ingestion	Apache Kafka, Apache Flume and Sqoop, Fiware Orion Context Broker
Data Storage	MongoDB [21], Apache HBase and Apache Hadoop Distributed File Systems (HDFS)
Data Processing Engines	Apache Spark and Apache Flink
Querying-Analytics-Visualization	Hive, Apache Spark MLlib, Apache Drill ,Apache Zeppelin, Mahout
Others (APIs, Reverse proxy coordination, Deployment)	Spring Framework[22] , Hadoop YARN and Apache Zookeeper, Traefik[23], Docker and Docker Swarm

In the Table 1 table is synthetized the main technology stack available in the architecture, most of them supported by the Apache Foundation [25]. Either of the technology has its advantages and disadvantages depending on the end, as example, regarding data collection, Apache Sqoop is specialized in RDBMS's like MySQL, Oracle, etc. into HBase, Hive or HDFS (or the other way back). While, on other hand, Apache Flume is good dealing with streaming sources that are continuously generated, such as Log files. So, each one of the technologies are important in a particular case, since we are dealing with different systems, not interoperable in many cases, in a complex factory environment.

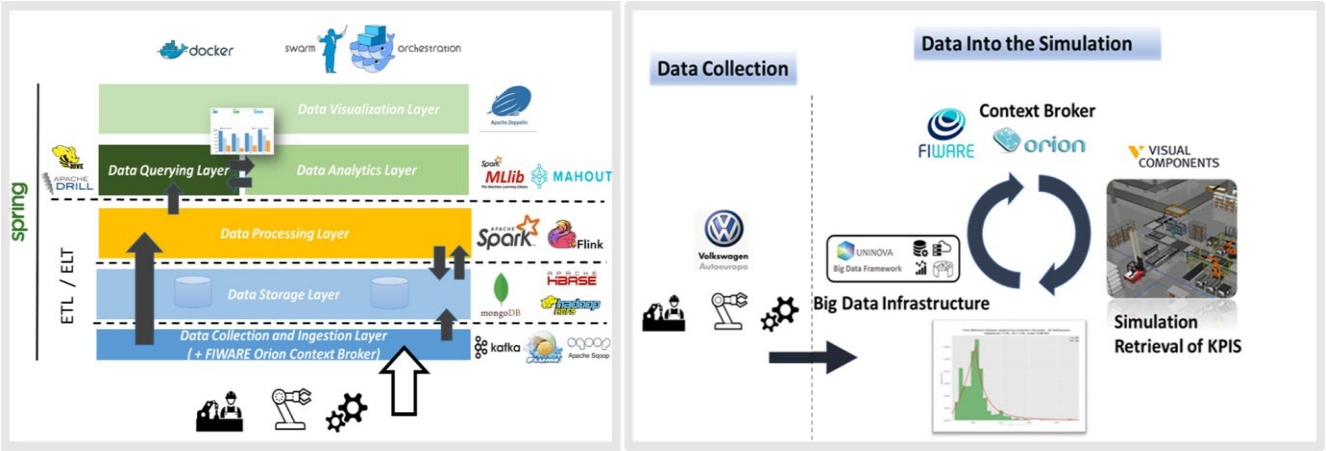

Figure 2: Reference Architecture (in the left) and a diagram showing how data feeds the 3D Digital Twin (in the right).

4.2 DEPLOYMENT: DOCKER AND DOCKER SWARM ORCHESTRATION

One of the key features of the approach presented here is how technologies are deployed and coordinated, namely to ease reproducibility, re-configurability, scalability, maintenance and support, as well as take advantage of technologies that require several nodes, in order to get the best performance. To do that, we use Docker Swarm. Docker is a virtualization software that utilize containers, which are isolated software packages ready to run. Docker Swarm is a clustering, managing and scheduling tool for Docker containers, deployed as managed as one virtual system and each computer node can run several containers with different software packages. After a Cluster is created, services (in this case the architecture´s technology stack) can be deployed to it using a Docker-compose file, with a description of how and where each "software package" must run, and how failures area handle. As example, we can have inside the Docker Swarm several smaller Clusters running different technologies, such as for data collection with Flume, data storage in HDFS, or distributed processing work with Apache Spark. Each with their masters and worker.

All the technologies that comprise the architecture can be defined in one Docker-compose file and deployed as a whole. Moreover, Docker Swarm handle new nodes and dropdowns, starting or restarting automatically services and jobs that were not complete, having in that sense self-management and easy scalability (taking advantage of the synchronization that the used Big Data technologies already do when deployed in several nodes, as for example, data-blocks replication in Hadoop). Figure 3, below, shows an example of a Docker Swarm Cluster running several services.

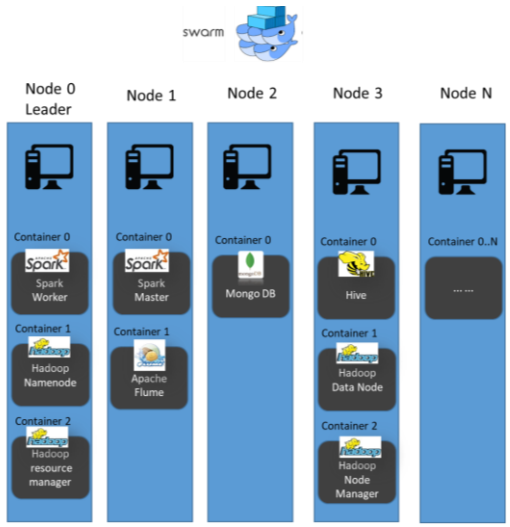

Figure 3: Example of technologies running in several nodes in a Docker Swarm Orchestration.

5. 3D DIGITAL TWIN

The 3D Digital Twin is achieved by associating data generated in the logistic process and a 3D Simulation in Visual Components framework, though the 3D visuals are not the priority, the simulation that has realistic layout and logistic movement of the factory. The Digital Twin, until now, was developed in two phases where different kinds of data is used to replicate the real operation: (1) statistical distributions of times the movements should occur, produced from Data Analytics applied on historical data, (2) ingest data as real-time to better emulate real world operations. Both using data previously acquired and processed by the Data infrastructure.

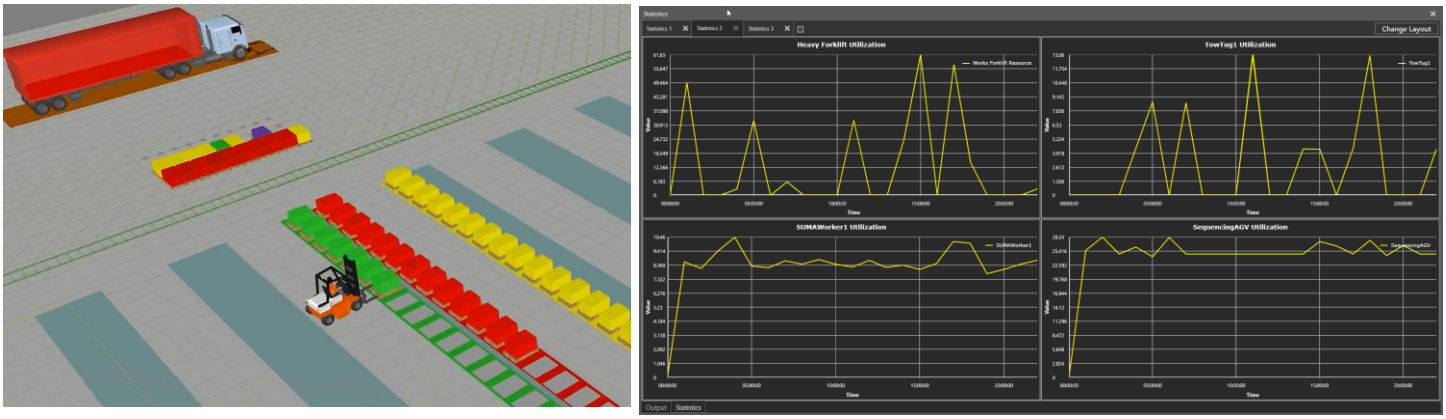

Figure 4: 3D Digital Twin: Warehousing image (in the left) and example of preliminary KPIs regarding the utilization of resources, such as the forklift utilization time (in the right).

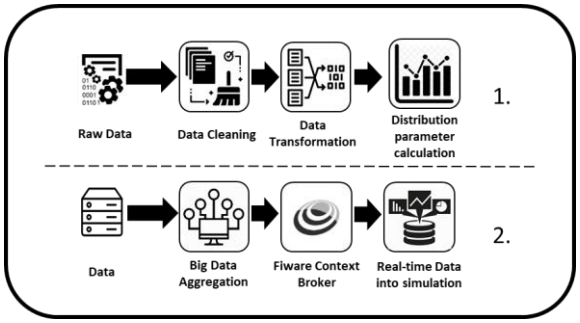

Figure 5: Digital Twin Analytics Level experiment phases.

A Simulation Controller, based on PySpark library (Apache Spark library for Python), and supported by the Data Infrastructure presented here, is responsible for the aggregation of the data of interest for each process (Receiving, Warehousing, Sequencing, Moving, Assembly), process it, and publishes to FIWARE Orion Context Broker, which manage the context information using MongoDB. From the other end, the VC simulation subscribes to entities (one for each process) and receives updates on their values, where each entity represents a timestamp and a movement to happen in the simulation, creating a buffer of action to perform. Since the simulation environment enables simulation speeds much higher than real-time, this data buffer is needed on the simulation environment's side and multiple data records are published at the same time as arrays. From the VC simulation is also possible to extract real-time analytics on the efficiency of different points of interest.

Until now , the focus was the Receiving and Warehousing processes, yet, the objective is to have the complete batteries operation working with real-data, so that is possible to test multiple logistic scenarios for future planning, such as increase the jobs per hour and see how the system´s behave, or change the sequencing area, getting back realistic feedback from the simulation. These kinds of tests, today, are only possible with by stopping production, which is extremely difficult to accomplish in a high production facility

6. CONCLUSIONS AND FUTURE WORK

This study tackles the challenge of how to process Big Data in an automotive car manufacturing scenario, working towards the creation of a Digital Twin for intra-logistics process planning to mimic the real-world logistics processes. Such Digital Twin can signify significant savings in time and production loss, since it can substitute many production stops needed to perform real tests in the production line, and by joining Data Analytics and KPIs, it can help identified where the efficiency can be majored.

In terms of architecture, it addresses the business and technical requirements and tries to aggregate the necessary technological stack, in one easy deployment with Docker Swarm, allowing distributed and parallel processing across computer nodes easily, automatic scalability, flexible and simplify management of the entire data infrastructure, since technologies is deployed, orchestrated and manage as one virtual system. It also follows BOOST4.0 orientation towards interoperability, namely for future full support of EIDS open source ecosystem that promise to connect and join factories and data service providers in the "Industrial Internet".

Concerning future work, the objective is to (i)better integrate the data into VC simulation, feeding more simulation processes with array data form the controller, either real-data or predictive data to fast-forward the simulation into the future and get KPIs, with the possibility to test some processes and have others automatically adapted to key variables (such as buffer values); (ii) have an User Interface for the controller, where the user can choose data wants to ingest into the simulation, start and stop the simulation, tweak variables (e.g. jobs per hour, AGVs velocity, initial conditions), and have some additional analytics that use also the feedback from the simulation KPIs.

ACKNOWLEDGEMENTS

Authors acknowledge the European Commission for its partial funding and the partners of BOOST4.0 research project - Agreement Number 780732.

REFERENCES

[1] Y. Yin, K. E. Stecke, and D. Li, "The evolution of production systems from Industry 2.0 through Industry 4.0," *Int. J. Prod. Res.*, vol. 56, no. 1–2, pp. 848–861, 2018.

[2] L. Da Xu, E. L. Xu, and L. Li, "Industry 4.0: state of the art and future trends," *Int. J. Prod. Res.*, vol. 7543, pp. 1–22, 2018.

[3] Dr. Henning Kagermann, "Recommendations for implementing the strategic," *Natl. Acad. Sci. Eng.*, no. April, 2013.

[4] R. Baheti and H. Gill, "Cyber-physical Systems," 2011.

[5] A. De Mauro, M. Greco, and M. Grimaldi, "What is big data? A consensual definition and a review of key research topics," in *INTERNATIONAL CONFERENCE ON INTEGRATED INFORMATION (IC-ININFO 2014): Proceedings of the 4th International Conference on Integrated Information*, 2014, vol. 1644, no. 1, p. 97–1De Mauro, A., Greco, M., Grimaldi, M. (2014).

[6] N. T. Tariq RS, "Big Data Challenges," *Comput. Eng. Inf. Technol.*, vol. 04, no. 03, 2015.

[7] Q. Qi and F. Tao, "Digital Twin and Big Data Towards Smart Manufacturing and Industry 4.0: 360 Degree Comparison," *IEEE Access*, vol. 6, pp. 3585–3593, 2018.

[8] R. Y. Zhong, X. Xu, E. Klotz, and S. T. Newman, "Intelligent Manufacturing in the Context of Industry 4.0: A Review," *Engineering*, vol. 3, no. 5, pp. 616–630, 2017.

[9] BOOST4.0 CONSORTIUM, "Boost 4.0 | Big Data for Factories," 2018. [Online]. Available: https://boost40.eu/. [Accessed: 05-Dec-2018].

[10] B. Otto *et al.*, "White paper: Industrial data space," 2016.

[11] Visual Components Oy, "Visual Components: 3D manufacturing simulation and visualization software - Design the factories of the future," 2019. [Online]. Available: https://www.visualcomponents.com/. [Accessed: 20-Mar-2019].

[12] Docker, "Swarm mode overview | Docker Documentation," 2018. [Online]. Available: https://docs.docker.com/engine/swarm/. [Accessed: 03-Jan-2019].

[13] K. Witkowski, "Internet of Things, Big Data, Industry 4.0 – Innovative Solutions in Logistics and Supply Chains Management," *Procedia Eng.*, vol. 182, pp. 763–769, Jan. 2017.

[14] J. Lee, B. Bagheri, and H. A. Kao, "A Cyber-Physical Systems architecture for Industry 4.0-based manufacturing systems," *Manuf. Lett.*, vol. 3, pp. 18–23, 2015.

[15] M. Nino, F. Saenz, J. M. Blanco, and A. Illarramendi, "Requirements for a big data capturing and integration architecture in a distributed manufacturing scenario," *IEEE Int. Conf. Ind. Informatics*, pp. 1326–1329, 2017.

[16] Y. Lu and X. Xu, "Cloud-based manufacturing equipment and big data analytics to enable on-demand manufacturing services," *Robot. Comput. Integr. Manuf.*, vol. 57, no. June 2018, pp. 92–102, 2018.

[17] M. Y. Santos *et al.*, "A Big Data system supporting Bosch Braga Industry 4.0 strategy," *Int. J. Inf. Manage.*, vol. 37, no. 6, pp. 750–760, 2017.

[18] F. Tao, J. Cheng, Q. Qi, M. Zhang, H. Zhang, and F. Sui, "Digital twin-driven product design, manufacturing and service with big data," *Int. J. Adv. Manuf. Technol.*, vol. 94, no. 9–12, pp. 3563–3576, Feb. 2018.

[19] C. Shearer *et al.*, "The CRIS-DM model: The New Blueprint for Data Mining," *J. Data Warehousing14*, vol. 5, no. 4, pp. 13–22, 2000.

[20] "Home - Fiware-Orion." [Online]. Available: https://fiware-orion.readthedocs.io/en/master/index.html. [Accessed: 15-Nov-2018].

[21] I. MongoDB, "Open Source Document Database | MongoDB," 2018. [Online]. Available: https://www.mongodb.com/. [Accessed: 15-Jan-2019].

[22] Pivotal Software, "Spring," 2019. [Online]. Available: https://spring.io/. [Accessed: 14-Jan-2019].

[23] Containous, "Traefik - The Cloud Native Edge Router," 2019. [Online]. Available: https://traefik.io/. [Accessed: 15-Jan-2019].