Technical Report

Big Data Needs and Challenges in Smart Manufacturing: An Industry-Academia Survey

Dietmar Winkler[1]
Alexander Korobeinykov[1]
Petr Novak[2]
Arndt Lüder[3]
Stefan Biffl[1]

[1] CDL-SQI, TU Wien, Austria
[first.last]@tuwien.ac.at

[2] Czech Institute of Informatics, Robotics, and Cybernetics,
Czech Technical University, Czech Republic
[first.last]@cvut.cz

[3] Otto-von-Guericke Universität Magdeburg, Germany
[first.last]@tuwien.ac.at

Technical Report No. CDL-SQI 2021-05
Issued: May 2021

Citation: D. Winkler, A. Korobeinykov, P. Novak, A. Lüder, S. Biffl “Big Data Needs and Challenges in Smart Manufacturing: An Industry-Academia Survey”, Technical Report CDL-SQI 2021-05, TU Wien, Vienna, Austria, May 2021, submitted to ETFA 2021 (under review)

Big Data Needs and Challenges in Smart Manufacturing: An Industry-Academia Survey

Dietmar Winkler*§, Alexander Korobeinykov*, Petr Novák‡, Arndt Lüder†**, Stefan Biffl§**
* Christian Doppler Laboratory for Security and Quality Improvement in the Production System Lifecycle,
§ Inst. of Information Systems Eng., TU Wien and ** CDP, Austria. Email: {firstname.lastname}@tuwien.ac.at
† Otto-von-Guericke University, Magdeburg, Germany. Email: arndt.lueder@ovgu.de
‡ Czech Institute of Informatics, Robotics, and Cybernetics, *Czech Technical University in Prague*,
Prague, Czech Republic. Email: {firstname.lastname}@cvut.cz

***Abstract*—The increasing availability of data in Smart Manufacturing opens new challenges and required capabilities in the area of big data in industry and academia. Various organizations have started initiatives to collect and analyse data in their individual contexts with specific goals, e.g., for monitoring, optimization, or decision support in order to reduce risks and costs in their manufacturing systems. However, the variety of available application areas require to focus on most promising activities. Therefore, we see the need for investigating common challenges and priorities in academia and industry from expert and management perspective to identify the state of the practice and promising application areas for driving future research directions. The goal of this paper is to report on an industry-academia survey to capture the current state of the art, required capabilities and priorities in the area of big data applications. Therefore, we conducted a survey in winter 2020/21 in industry and academia. We received 22 responses from different application domains highlighting the need for supporting (a) fault detection and (b) fault classification based on (c) historical and (d) real-time data analysis concepts. Therefore, the survey results reveals current and upcoming challenges in big data applications, such as defect handling based on historical and real-time data.**

***Index Terms*—Smart Manufacturing, Big Data Application, State of the Practice, Required Capabilities, Survey.**

I. Introduction

Industry 4.0 initiatives aim at addressing business demands for flexible production in terms of product variants and volume [2] supported by increased digitalization [8]. In this context, *smart manufacturing* refers to a high level of adaptability, flexibility, the ability to respond to rapid changes, and digital information technology [30]. The increasing availability of data in context of smart manufacturing opens new challenges and expected capabilities for data analysis in industry and academia [16]. Various organizations have started initiatives to collect and analyse data in their individual contexts with specific goals, e.g., for monitoring, optimization, or decision support. For example, GAiA-X[1] aims at developing common requirements for a European data infrastructure as foundation for data analysis. Zheng *et al.*report on a conceptual framework, scenarios, and future perspectives in smart manufacturing for *Industry 4.0* [30]. Lu *et al.* [11] review standards in smart manufacturing process and system automation.

[1]GAIA-X: www.data-infrastructure.eu

Big Data concepts and technologies provide corner-stones for smart manufacturing, e.g., logging, data processing, and analytics [28]. Given a huge range of different research and application initiatives, there is a need for prioritization of *Big Data* analysis approaches to focus on most critical application areas in industry and academia. The identification of current needs, expected capabilities of *Big Data* applications and their prioritization can help to better support the definition of strategic steps for driving improvement initiatives in industry and academia.

Therefore, we see the need for investigating needs, challenges and priorities from expert and management perspective to identify the state of the practice in organizations as foundation for identifying promising application areas and future trends in industry and academia. The **goal of the study is to identify the state of the practice, related challenges, expected capabilities, and perceived priorities of *Big Data* in smart manufacturing in academia and industry.**

Therefore, we conducted a survey from December 2020 to February 2021 in industry and academia to identify the state of the practice in *Big Data* applications in context of challenges, expected capabilities and perceived priorities in organizations. Most of the respondents are experts and managers in positions of organizations that have a good overview on innovation topics and projects in context of smart manufacturing, *Big Data*, and *Data Analytics*. Therefore, we can expect high-quality data from this survey.

We received 22 responses from different application domains, e,g., machine manufacturing, automotive, aerospace & defence, and logistics. The results highlight the need for solutions that support (a) fault detection and (b) fault classification based on (c) historical and (d) real-time data - perceived as topics with highest priorities by respondents. Addressing these prioritized needs and expected capabilities can help decreasing risks and cost in current and future smart manufacturing systems. The results go beyond current research reports by taking into considerations perceived priorities from industry and academia from expert and management perspective.

The remainder of this paper is structured as follows: Section II summarizes related work on smart manufacturing and big data analytics. We present our research questions in Section III. Section IV presents the survey structure. Section V

presents the results. We discuss the results in Section VI. Finally, Section VII concludes the paper and proposes future work.

II. Related Work

This section summarizes related work on *Smart Manufacturing*, *Big Data Analytics*, and the application of *Big Data* in *Smart Manufacturing*.

A. Smart Manufacturing

Smart manufacturing is related to a new generation of cyber-physical production systems featured with data connectivity and artificial intelligence [25]. Numerous definitions and understanding of this term can be found, e.g., in [24]. According to the material of U.S. National Institute of Standards and Technology (NIST), smart manufacturing systems are understood as *"fully-integrated, collaborative manufacturing systems that respond in real time to meet changing demands and conditions in the factory, in the supply network, and in customer needs"*[2]. From the technology point of view, smart manufacturing systems provide additional advanced features covering platforms for *Industrial Internet of Things*, big data processing of production data and *Cloud Computing*, additive manufacturing, computer-integrated manufacturing, advanced robotics, and artificial/machine intelligence [24].

The term *smart manufacturing* is, especially in Europe, frequently related to the term *Industry 4.0* [26]. Industry 4.0 is a vision/strategy on how to shift traditional industrial production systems towards modern flexible systems, benefiting from advances in the area of Artificial Intelligence (AI). According to Etz *et al.*, *"smart manufacturing is realizing the idea and potential of Industry 4.0 in reality"* [7].

Transforming a traditional production facility towards smart manufacturing is a long-term process. A roadmap for this shift is proposed in [23]. It consists of six pillars, called six gears: (i) Strategy, (ii) Connectivity, (iii) Integration, (iv) Data analytics, (v) Artificial Intelligence, and (vi) Scalability [23]. Lu *et al.* [11] provide a review of existing industrial standards that are capable to support smart manufacturing approaches and system automation. Data acquisition and data processing play fundamental roles in smart manufacturing [4]. Nagorny *et al.* [16] provide a review on big data analysis in smart manufacturing. Smart manufacturing is frequently supported by the concept of digital twins and the use of big data in conjunction with a digital twin is addressed in [21].

Given the variety of topics in smart manufacturing, current needs and expected capabilities in context of data analytics can help to identify most relevant aspects to be addressed in industry and academia.

B. Big Data Analytics

The amount of machine-generated data is growing fast, especially in the manufacturing domain [31]. The datasets become so enormously large and complex that conventional database systems cannot process them within the desired time [12]. Data of this magnitude are referred to as Big Data [27].

The definition of *Big Data* frequently relies on the so-called *V characteristics* which have changed over the years as more knowledge was gathered and a deeper understanding of the nature of Big Data has formed. In 2014, the definition contained only the "**3V**" characteristics of the data - *volume*, *variety*, and *velocity* [3]. In 2015, the definition was extended with *veracity* [13], in 2017 with *value* [25], emphasizing the importance of the qualitative data characteristics. In 2018, "**10V**" characteristics [21], further included *vision, volatility, verification, validation*, and *variability*. We expect the number of Big Data "**V**" characteristics to grow further as the community expertise grows and the domain gets more mature.

A traditional approach to organizing and processing large amounts of data is a *Data Warehouse* with processes that build on batch processing and data pipelines organized according to the *Extract-Transform-Load* (ETL) or *Extract-Load-Transform* (ELT) principle [1]. With increasing data velocity, the value of data decays faster. Traditional warehouse systems and the employed batch-oriented processing models cannot provide the expected results within the posed latency constraints [29]. As a consequence, in the smart manufacturing, real-time processing of collected data is needed to enable fast and efficient results for analysis and decision making.

Therefore, the need for *stream data processing* becomes obvious in many industries [18], in particular, for manufacturing with high-velocity processes and strong dependencies between process steps. For example a continuous flow of data, provided by the smart manufacturing system, needs to be processed to receive fast analysis results for taking efficient actions. *Lambda and Kappa architectures* for data analysis are applied to address the increasing data velocity and were already successfully realized in varied contexts [20] [10].

Based on the classification of different *Big Data* characteristics, in the survey we focus on the most critical characteristics related to smart manufacturing.

C. Big Data in Smart Manufacturing

Big Data concepts have been applied in manufacturing to improve data management and to enable advanced analytical functionalities, such as machine learning and Artificial Intelligence. Windmann *et al.* describe the application of *Big Data Analytics* for anomaly detection and for PCA-based anomaly prediction in different manufacturing scenarios [28]. Tao *et al.*outline the cloud-based manufacturing data analytics system designed to support data-driven product quality control and smart equipment maintenance [25]. Moyne *et al.* proposed the predictive maintenance system for a semiconductor factory, based on product and equipment data [15]. Big Data concepts were used to improve maintenance practices by identifying patterns in the data [15]. Similar research was conducted by O'Donovan *et al.* by designing a manufacturing data processing pipeline to improve equipment maintenance that utilized cloud services for data aggregation and analytics [17].

[2]NIST: www.nist.gov/programs-projects/smart-manufacturing-operations-planning-and-control-program

Lee *et al.* report the successful implementation of the Big Data analytics system in the small to medium-sized manufacturing environment [9]. The cloud-based analytics system aimed at supporting process monitoring and provided basic product defect prediction. While evaluating cloud options for manufacturing data management many aspects, such as data security, data governance and related legal issues, could impede the further development of smart manufacturing principles. The GAIA-X project aims at tackling those issues for European enterprises. The platform aims to support European data initiatives with a shareable solution for data management on the cloud. Such initiatives can relax the above-mentioned requirements, such as a high volume and the availiability of data for processing, and ease the the implementation risk of the Big Data concepts for smart manufacturing.

As the applications and initiatives in the domain are pretty dispersed at the moment, there is an apparent need to better understand the current state of the practice, plans and priorities for the future development from practitioners in the industry and from academic researchers closely working on the topic. An online survey was used to capture needs and expected capabilities including the perceived prioritization from an expert and manager perspective in the field of smart manufacturing.

III. Research Questions

Based on the goal of this paper and related work, we derived the following research questions.

RQ1. What are the current challenges in Big Data and Smart Manufacturing? Based on reported challenges, e.g., in [12] [31], the goal of the survey is to identify and assess current needs from expert and management perspective in the smart manufacturing context in industry and academia. Main results identify to what extent experts and managers see identified challenges in their smart manufacturing context.

RQ2. What are the estimated priorities of Smart Manufacturing capabilities? Because of the wide range of various topics in the *Big Data* area in smart manufacturing, the question focuses on the perceived importance of experts and managers in terms of prioritization.

*RQ3. What are the estimated priorities of Big Data functionalities in Smart Manufacturing?*To explore functions in context of data analytics and related priorities that a *Big Data* application in smart manufacturing should provide, we derive the third research question.

RQ4. What is the readiness of organizations to use cloud infrastructures for smart manufacturing? Finally, the growing number of data and cloud storage options, this question focuses on the readiness of organization to apply cloud options for data storage and analytic purposes. Therefore, the survey addresses concerns on the readiness of organizations for cloud infrastructures.

To address the research question and to collect experiences from experts and managers in the smart manufacturing domain, we conducted an online survey ([6], [14], [19]) following empirical software engineering standards on survey design and data analysis.

IV. Survey Design

This Section summarizes the survey procedure and the setup of the survey questionnaire based on empirical software engineering survey best practices for the design and data analysis [6], [14], [19], [22].

A. Survey Procedure

Physical meetings or distribution and processing of physical materials was not possible due to the COVID-19 imposed constraints in the period of surveying. Therefore, the study plan focused on conducting an online expert survey.

Survey Process. Figure 1 describes the main steps of the survey: (A) *Survey Design* (see Section IV-B); (B) *Survey Execution*, We use Google.Forms[3] for designing, conducting, and analyzing the survey; and (C) *Data Analysis* that focuses on descriptive statistics of the survey responses, provided by Google.forms. In addition, we used a spreadsheet solution for analyzing descriptive statistics. We applied Boxplots, Bar Charts, Pie Charts, and descriptive statistics for reporting.

Participants. Candidate participants were selected based on existing collaborators of the involved research groups (i.e., the authors) and include industrial partners and academic groups from other universities and research organizations. The candidate participants in this selection are situated in Europe, which justifies the addition of context-dependent questions in the cloud section.

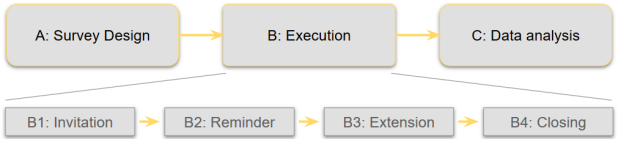

Fig. 1. Survey Process Overview.

The *Survey Execution* phase (see Label B in Fig 1) consists of four process steps:

(B1) Invitation. The invitations to selected participant groups were sent out in December 2020. The invitation includes a suggestion to share this survey further. The end date was communicated to the participants and additionally stated in the description of a distributed survey.

(B2) Reminder. In January 2021, a reminder was sent out to all the email addresses saved in the list during the survey distribution. This measure is required, as in many cases people tend to postpone tasks that are not of a high priority and can be accomplished later on. This is especially relevant in case of the vacation season.

(B3) Extension. Due to low response rate, the survey duration was extended till the end of February 2021. In addition, to increase the number of responses, the survey was distributed in social networks such as LinkedIn [4] and ResearchGate [5].

(B4) Closing. On February 28th, 2021 the survey was closed for the initial survey round. We used Google.forms

[3]Google.Forms: forms.google.com
[4]LinkedIn: www.linkedin.com
[5]ReserachGate: www.researchgate.net

and a spreadsheet solution for analyzing the responses and for creating statistics.

B. Survey Design

The survey is constructed to source information about particular needs of a *smart manufacturing* domain. Survey questions include (a) *Background* of the respondents; (b) *Smart Manufacturing and Data Analytics* including needs and challenges, prioritization, and expected capabilities; and (c) *Cloud Options for Smart Manufacturing* with focus on readiness for the cloud. Finally, (d) the optional part covering *Feedback and Contact information* complete the survey. We use this personal data of respondents for getting in contact for clarifying questions, conducting in-depth interviews, and for providing additional information. The survey implementation can be found online[6].

The study plain aimed at a data collection effort for respondents of up to *10* minutes. Most questions were designed as multiple choice complemented by open text questions to add additional information. The survey did not require any authentication and could be answered anonymously. As an introduction, the aim and short description of the survey were provided. Along with the other process rules, the survey instructions explicitly mentioned the scientific purpose of the research, i.e., data gathered will not be used for commercial purposes. The survey is divided into the following four parts.

(a) Background and Respondent Categorization. The first part contains questions on respondent background, their field of work, position, working experience, and the typical size of the projects that are conducted in their company.

(b) Manufacturing and Data Analytics. The second part focuses on manufacturing capabilities and capabilities for the data analytics functionality. Respondents are asked to estimate the importance of specific and pre-defined manufacturing capabilities and data processing paradigms on a 5-point Likert scale. Furthermore, an optional free-text field can be used to add additional comments. Respondents were also asked to evaluate currently existing needs/challenges in their organization. The pre-defined set of needs and challenges have been derived from literature.

(c) Cloud options for Smart Manufacturing. The third part is about the possibilities and readiness of the industry to use cloud solutions to store and process the manufacturing data. Data analysis can help to better understand the attitude towards cloud technologies and to determine how reasonable it is to propose a solution that involves cloud resources. The section includes region-specific questions for Europe: use of the regional EU cloud service.

(d) Feedback and Contact Information. In this part, respondents can provide their contact data to receive feedback or participate in future research activities.

V. Survey Results

The data analysis was based on the valid answers coming from *22* respondents. Most of the respondents were directly invited, a small number of respondents knew about the survey from open communities.

[6]Survey: https://forms.gle/sA9MGBz4dBxuzshcA

A. Applications Domains and Experience

The participants mostly represent the automotive, machine manufacturing, and academia domains (see Table I and Figure 2).

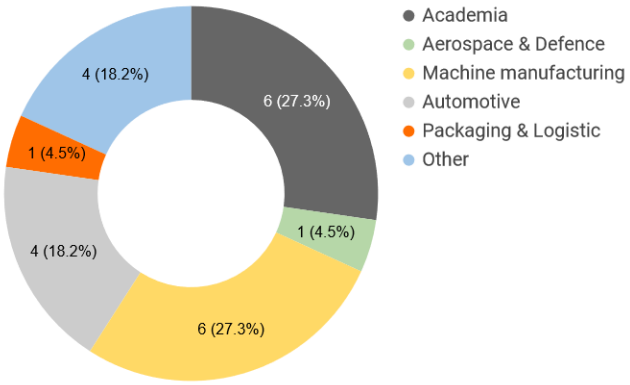

Fig. 2. Distribution of survey participants by domain (cf. Table I).

TABLE I
DISTRIBUTION OF THE SURVEY PARTICIPANTS BY INDUSTRY.

Domain	Number of respondents (%)
Academia	6 (27.3%)
Aerospace & Defence	1 (4.5%)
Machine manufacturing	6 (27.3%)
Automotive	4 (18.2%)
Packaging & Logistic	1 (4.5%)
Other	4 (18.2%)
Total	22 (100%)

The roles of the respondents indicate that they should possess significant experience in the field of occupation and due to mostly managerial positions, should be well aware of the current challenges and future development vectors of their companies. We collected the position as free text to enable reporting on the current position within the organization. The majority of the reported roles can be summarized as engineering *management roles*, including Senior Manager in Innovation, R&D Team Lead, Project Leader, CIO, Project Manager, Head of Innovation and software Development, Center Manager, Applied Artificial Intelligence Unit Director, Head of ICT & Automation Department, Head of Business Unit, and Manager Analytics and AI. *Non-management roles* include Researcher, Application Engineer, Technology Consultant, Project Engineer, and Data Analyst. Most of the participants have reasonable experience in the smart manufacturing domain, with *50%* of the respondents involved in the topic for over *5* years. Taking into account that most of the respondents are either middle or C-level managers (positions in the companies that usually have a good overview), we can expect high-quality data from this survey.

Following the respondent's identification, it was important to identify the size and longevity of typical projects. This

information is useful to determine whether the companies invest in big, long-running initiatives or rather concentrate on short projects where the results can be expected sooner. The projects mostly conducted in the companies of the respondents are medium-size projects with *3-8* person-years scope (*50%*). Small and large sizes were reported with comparable frequencies of *22.7%* and *27.3%*, respectively.

Understanding the typical company's project sizes is important for designing a Big Data Analytics system and planning its integration into an existing manufacturing landscape. The companies of the respondents mostly engage into small and medium size projects. Therefore, a modular architecture for a Big Data Analytics system should be pursued to allow an iterative implementation and integration approach.

B. Smart manufacturing and Data Analytics

The section contained questions and free-text comments. This allowed the respondents not only to choose from the options provided but also to add new points or company-specific inputs from their practice. The questions that contained 2-dimensional answers (to evaluate relevance on the scale from *1* to *5* for a variety of features) are presented in the form of box plots to provide a better overview on the results.

First, respondents were asked to identify current challenges in their company. The challenges have been derived based on related work and from literature. The relevance data for the challenges outlined in the survey can be seen in Table II and Figure 3. Challenge "CH2: high data velocity", as well as "CH1: high data volumes", received a very consistently high score from most of the participants. Challenges "CH4: data transformation" and "CH5: data analysability" reflected a steady progression and showed an increase in the number of votes on the scale from mark *1* to mark *5*. On the other side, challenges such as "CH8: lack of data insights" and "CH7: manual data processing" followed the normal statistical distribution and received the most number of votes in the middle of the point scale.

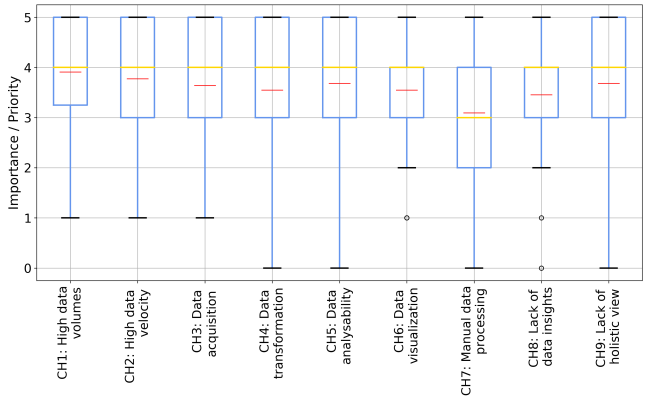

Fig. 3. Importance/priority of challenges (cf. Table II). Yellow/red lines mark median/mean values.

When asked which of the challenges are currently being addressed in the companies of the respondents, the following

TABLE II
DESCRIPTIVE STATISTICS FOR IMPORTANCE OF CHALLENGES (CF. FIGURE 3).

Challenge	mean	median	std.dev.
CH1: High data volumes	3.91	4.00	1.15
CH2: High data velocity	3.77	4.00	1.23
CH3: Data acquisition	3.63	4.00	1.49
CH4: Data transformation	3.54	4.00	1.43
CH5: Data analysability	3.68	4.00	1.46
CH6: Data visualization	3.54	4.00	1.18
CH7: Manual data processing	3.09	3.00	1.31
CH8: Lack of data insights	3.45	4.00	1.29
CH9: Lack of holistic view	3.68	4.00	1.46

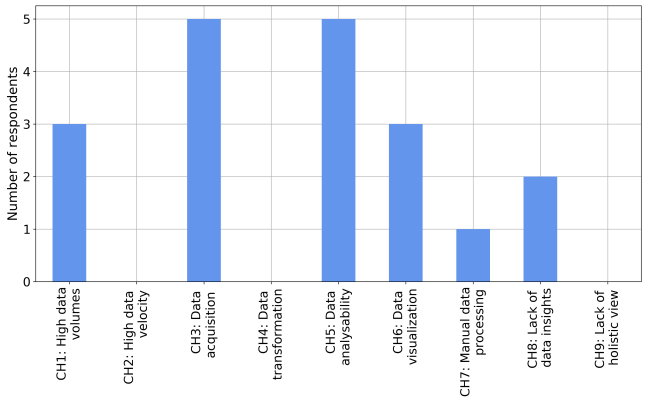

Fig. 4. Challenges currently addressed in the companies of the respondents.

results emerged: the most addressed challenges currently are the "CH3: data acquisition" and "CH5: data analysability" followed by "CH1: high data volumes" and "CH6: data visualization" challenges (see Figure 4). Some of the challenges, such as "CH2: high data velocity", "CH4: data transformation", and "CH9: lack of holistic view", were reported as not being addressed in the companies of the respondents.

The expected capabilities of the smart manufacturing system were listed for relevance evaluation as well. A list for these capabilities was formed based on the research of Moyne *et al.* [15] and reviewed together with the industry partners. The relevance data for the smart manufacturing system's capabilities can be seen in Table III and Figure 5. Not surprisingly, the "C1: fault detection" and the "C2: fault classification" lead the list. These two capabilities received most of their votes in the higher range of points scale (mostly *4* and *5* points). Capabilities "C5: statistical process control" and "C9: yield prediction" followed the normal statistical distribution with most of the votes contained in the middle of the points scale. Note that capability "C8: predictive scheduling" also received a significant amount of votes in the middle and high range but due to some responses marking this capability as not applicable in their environment the overall score is still lower than other capabilities.

Regarding the functionalities of the data analytics system to support the smart manufacturing environment, most of the respondents emphasize functionalities "F1: Historical data analysis" and "F2: real-time data analysis" (see Table IV

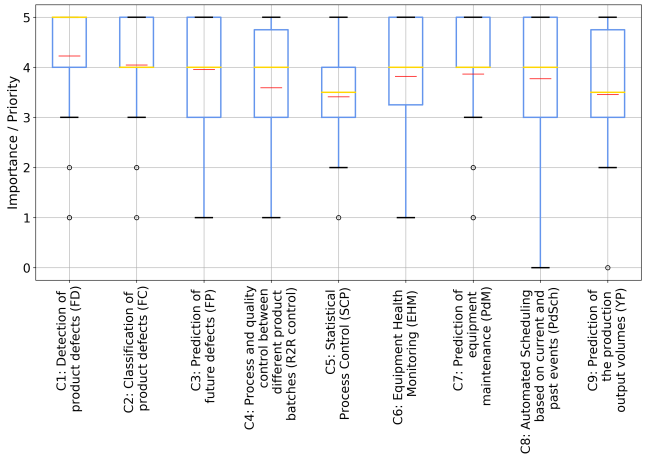

Fig. 5. Importance/priority of smart manufacturing capabilities (cf. Table III). Yellow/red lines mark median/mean values.

TABLE III
DESCRIPTIVE STATISTICS FOR IMPORTANCE OF SMART MANUFACTURING CAPABILITIES (CF. FIGURE 5).

Capability	mean	median	std.dev
C1: Fault Detection (FD)	4.22	5.00	1.31
C2: Fault Classification (FC)	4.05	4.00	1.17
C3: Fault Prediction (FP)	3.95	4.00	1.17
C4: Run-2-Run control (R2R)	3.59	4.00	1.22
C5: Statistical Process Control (SPC)	3.41	3.50	0.96
C6: Equipment Health Monitoring (EHM)	3.81	4.00	1.41
C7: Predictive Maintenance (PdM)	3.86	4.00	1.28
C8: Predictive Scheduling (PdSch)	3.77	4.00	1.48
C9: Yield Prediction (YP)	3.45	3.50	1.34

and Figure 6). Once again, votes for these functionalities were mostly gathered in the higher range of points scale (mostly *4* and *5* points). The responses for functionalities, such as "F3: data mining", "F5: data transformation", and "F7: data storage" followed the normal statistical distribution and contained most of the votes in the middle of the points scale. Functionality "F4: data visualization" appeared to be a strongly expected system's capability as well, having most of the votes in the higher range of the points scale. From all of the presented options, only functionality "F8: time-series analysis" was reported to be not applicable to one of the respondent's use cases.

TABLE IV
DESCRIPTIVE TATISTICS FOR IMPORTANCE OF BIG DATA ANALYTICS FUNCTIONALITIES (CF. FIGURE 6).

Functionality	mean	median	std.dev.
F1: Historical data analysis	4.00	4.00	1.31
F2: Real-time data analysis	4.13	5.00	1.24
F3: Data mining	3.50	3.50	1.05
F4: Data visualization	3.77	4.00	1.37
F5: Data transformation	3.59	4.00	1.14
F6: Data transfer	3.41	4.00	1.40
F7: Data storage	3.22	3.00	1.23
F8: Time series analysis	3.68	4.00	1.39
F9: Multi-variate analysis	3.77	4.00	1.23

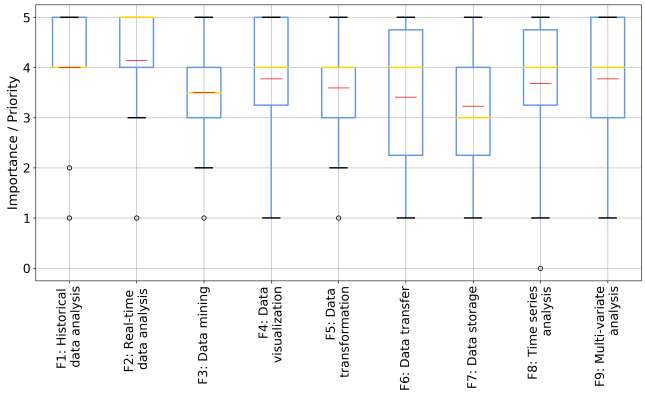

Fig. 6. Importance/priority of Big Data Analytics functionalities (cf. Table IV). Yellow/red lines mark median/mean values.

C. Cloud Options for smart manufacturing

Data reflecting the position of the respondents on cloud utilization in smart manufacturing are quite interesting. Most respondents (*81.8%*) indicate it possible to move parts of manufacturing data processing and storage to the cloud. *18.2%* of the respondents suggested, at least in theory, it to be possible to move all parts of manufacturing data processing and storage to the cloud. No respondent indicated cloud services not usable for data processing and storage within their smart manufacturing scenarios.

Almost all respondents (*95.5%*) reported that it is possible for their organization to move parts of the data storage and/or processing to the cloud. A large majority of the survey participants (*95.2%*) are ready to use European cloud services for that purposes. This means that the Big Data platforms utilizing cloud solutions can be practically introduced into the existing or planned smart manufacturing environment of the companies. The result emphasizes the assumptions about the possible role of cloud services made by other researchers, e.g., Moyne *et al.* [15] and Qi *et al.* [21]).

66.7% of respondents were aware of the existence of the GAIA-X project. Among those, industry and academia representatives have an equal share. However, many respondents from industry were not aware of the GAIA-X project.

VI. DISCUSSION AND LIMITATIONS

This section discusses the results by research question and limitations of the research.

RQ1. What are the current challenges in big data and smart manufacturing? Besides the challenges explicitly mentioned in the survey, the respondents were asked to provide any challenges they face as a free text. Only half of the respondents provided such challenges. Among the provided existing challenges are the following: "data protection", "data security", "lack of data quality", "lack of interoperability", "lack of scalability", "data heterogeneity", "the applicability of AI and machine learning", "price of the technologies required", and the "availability of solid and unified data architectures". A relatively high number of additionally provided challenges

suggests that their number and variety in the real setup is even bigger. This aligns well with the research results of LV *et al.* [12] and Chen *et al.* [5], who suggest that not only the number of common not well-addressed challenges are high, but the number of the challenges unique to every single situation should not be underestimated.

RQ2. What are the estimated priorities of Smart Manufacturing capabilities? The respondents were also asked to present additional capabilities of the SM system that were not on the above-mentioned list. Only *6* responses were submitted in this section one of them stating that no additional capabilities are required. Among the other responses to this question, the following topics were present: autonomous optimal control of machinery, Digital Twins, workforce management, maintenance recommendations and optimization analytics (mentioned twice), visualization of the process, and the data related to the above-mentioned capabilities (i.e. dashboarding). The low amount of the additional capabilities might have different meanings - either most of the required capabilities are presented in the list or the respondents do not have a clear vision of which capabilities they require as well.

RQ3. What are the estimated priorities of Big Data functionalities in Smart Manufacturing? Only *2* responses were received to the question of additional data analytics system functionalities missing from the provided list. Those were the "interoperability" and "pattern recognition" features coupled with the recommendation functionality available on steps of the data analytics process.

RQ4. What is the readiness of organizations to use cloud infrastructures for smart manufacturing? Traditional approaches for shop-floor cyber-security have been based on disconnecting low-level machinery from the public Internet. For many years, industrial companies have been reluctant to store or process production data outside the production site. Nevertheless, surprisingly high number of respondents agreed with the possibility to utilize cloud technologies for data storage or processing, which we perceive as a paradigm shift. Moreover, this shift can foster AI methods for massive manufacturing data processing, because computation power of cloud technologies can be easily scaled up for doing more complex analyses, compared to local databases close to industrial machines or server rooms onsite.

Limitations. In this survey, we focused on the identification of the state of the practice, challenges, expected capabilities, and perceived priorities of Big Data concepts in smart manufacturing. However, the survey contains a set of limitations:

Construct validity. The selection of the questions is a threat to validity with any survey study. Questions comprising this concrete survey aimed at the identification of current challenges in the research and industry and the most important vectors of future development of smart manufacturing concepts. To increase *construct validity* of the survey, apart from the respondent identification and feedback sections, the questions were created based on the capabilities and functionalities outlined in literature, e.g., [12], [31].

Some of the questions contained a predefined list of options, which limited the answers of the respondents, potentially creating a situation where respondents cannot find their actual answers represented. Therefore, the survey provided free text answer options to give users the possibility to add their own inputs to mitigate the threat of *limited levels of construct*. Note that the pre-defined answer set is based on literature (e.g., [12], [31]), complemented by an optional free-text answer option for possible extensions.

External validity. Bias may be introduced from respondent group selection. The invitations for participation were sent to the contacts from the research circles and industry partners that were known to the authors of this paper p from previous collaborations. Such collaborations usually happen around some common challenges and research topics. Therefore, this approach tends to increase the likelihood of respondents sharing similar views on the topic of *Big Data Analytics* and *smart manufacturing*, and pursue similar interests within a field. This introduces *threats to external validity*, which could be mitigated by aiming at a more diverse participant sample in future empirical studies.

VII. Conclusion and Future Work

The growing availability of data in the area of *Smart Manufacturing* opens new challenges for *Big Data* in organizations in academia and industry. Furthermore, organizations typically expect certain capabilities and functionalities from *Big Data* analytics in *Smart Manufacturing*.

In this paper we reported on a survey that aims at identifying challenges, expected capabilities and functionalities including their priorities perceived by organizations. In the survey, we received 22 responses from five different application domains highlighting the need for supporting (a) fault detection and (b) fault classification based on (c) historical and (d) real-time data analysis concepts. The results of this survey reveals current and upcoming challenges in big data applications. Note that the results are based on experts and managers in the field that have an overview on innovation and trends within their organizations. Finally, the outcome of the survey addresses the option to emphasize cloud solutions that can be practically introduced into the existing or planned smart manufacturing environment of the companies.

Future Work will include the repetition of the study, aiming at a larger and more diverse academic and industry context. Furthermore, the survey questions will be extended to gain more detailed insights into challenges, expected capabilities and functionalities, and priorities. Finally, based on the results of the survey, the goal is to derive promising research directions to drive research in academia and industry in context of *Big Data* in *Smart Manufacturing*.

Acknowledgment

The financial support by the Christian Doppler Research Association, the Austrian Federal Ministry for Digital & Economic Affairs and the National Foundation for Research, Technology and Development is gratefully acknowledged. The

competence center CDP is funded within the framework of COMET – Competence Centers for Excellent Technologies by BMVIT, BMDW, and the federal state of Vienna, managed by the FFG. This result was also funded by Ministry of Education, Youth and Sport of the Czech Republic within the project Cluster 4.0, reg. number CZ.02.1.01/0.0/0.0/16_026/0008432 and the RICAIP project funded by the Horizon 2020 research and innovation programme under grant agreement No. 857306.

REFERENCES

[1] Umar Ahsan and Abdul Bais. A review on big data analysis and internet of things. *IEEE 13th International Conference on Mobile Ad Hoc and Sensor Systems*, 2013. DOI: 10.1109/MASS.2016.38.

[2] Thomas Bauernhansl, Michael Ten Hompel, and Birgit Vogel-Heuser. *Industrie 4.0 in produktion, automatisierung und logistik: Anwendung-Technologien-Migration*. Springer, 2014.

[3] Ruben Casado and Muhammad Younas. Emerging trends and technologies in big data processing. *Concurrency and Computation: Practice and experience, Wiley Online Library (wileyonlinelibrary.com)*, October 01, 2014. DOI: 10.1002/cpe.3398.

[4] José María Cavanillas, Edward Curry, and Wolfgang Wahlster. The big data value opportunity. In José María Cavanillas, Edward Curry, and Wolfgang Wahlster, editors, *New Horizons for a Data-Driven Economy - A Roadmap for Usage and Exploitation of Big Data in Europe*, pages 3–11. Springer, 2016.

[5] Baotong Chen, Jiafu Wan, Lei Shu andPeng Li, Mithun Mukherjee, and Boxing Yin. Smart factory of industry 4.0: Key technologies, application case, and challenges. *IEEE Open access journal*, December 14, 2017. DOI: 10.1109/ACCESS.2017.2783682.

[6] Marcus Ciolkowski, Oliver Laitenberger, Sira Vegas, and Stefan Biffl. Practical experiences in the design and conduct of surveys in empirical software engineering. In *Empirical methods and studies in software engineering*, pages 104–128. Springer, 2003.

[7] Dieter Etz, Thomas Frühwirth, and Wolfgang Kastner. Flexible safety systems for smart manufacturing. In *2020 25th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, volume 1, pages 1123–1126, 2020.

[8] Heiner Lasi, Peter Fettke, Hans-Georg Kemper, Thomas Feld, and Michael Hoffmann. Industry 4.0. *Business & information systems engineering*, 6(4):239–242, 2014.

[9] Ju Yeon Lee, Joo Seong Yoon, and Bo-Hyun Kim. A big data analytics platform for smart factories in small and medium-sized manufacturing enterprises: An empirical case study of a die casting factory. *International Journal of Precision Engineering and Manufacturing Vol. 18*, October 2017. DOI: 10.1007/s12541-017-0161-x.

[10] Jimmy Lin. The lambda and the kappa. *IEEE Internet Computing*, 21(5):60–66, 2017.

[11] Yuqian Lu, Xun Xu, and Lihui Wang. Smart manufacturing process and system automation – a critical review of the standards and envisioned scenarios. *Journal of Manufacturing Systems*, 56:312–325, 2020.

[12] Zhihan Lv, Houbing Song, Pablo Basanta-Val, Anthony Steed, and Minho Jo. Next-generation big data analytics: State of the art, challenges, and future research topics. *IEEE Transactions on Industrial Informatics, Volume 13, Nn. 4*, August 2017. DOI: 10.1109/TII.2017.2650204.

[13] Nazim H. Madhavji, Andriy Miranskyy, and Kostas Kontogiannis. Big picture of big data software engineering. *IEEE/ACM 1st International Workshop on Big Data Software Engineering*, 2015. DOI: 10.1109/BIGDSE.2015.10.

[16] Kevin Nagorny, Pedro Lima-Monteiro, Jose Barata, and Armando Walter Colombo. Big data analysis in smart manufacturing: A review. *International Journal of Communications, Network and System Sciences*, 10(3):31–58, 2017.

[14] Jefferson Seide Molléri, Kai Petersen, and Emilia Mendes. Survey guidelines in software engineering: An annotated review. In *Proceedings of the 10th ACM/IEEE international symposium on empirical software engineering and measurement*, pages 1–6, 2016.

[15] James Moyne and Jimmy Iskandar. Bid data analytics for smart manufacturing: Case studies in semiconductor manufacturing. *MDPI Processes*, July 12, 2017. DOI: 10.3390/pr5030039.

[17] Peter O'Donovan, Kevin Leahy, Ken Bruton, and Dominic T. J. O' Sullivan. An industrial big data pipeline for data-driven analytics maintenance applications in large-scale smart manufacturing facilities. *Journal of Big Data, 2:25*, 2015. DOI 10.1186/s40537-015-0034-z.

[18] Gautam Pal, Gangmin Li, and Katie Atkinson. Multi-agent big-data lambda architecture model for e-commerce analytics. *MDPI, Data*, December 2018. DOI: 10.3390/data3040058.

[19] Teade Punter, Marcus Ciolkowski, Bernd Freimut, and Isabel John. Conducting on-line surveys in software engineering. In *2003 International Symposium on Empirical Software Engineering, 2003. ISESE 2003. Proceedings.*, pages 80–88. IEEE, 2003.

[20] Pekka Pääkkönen and Daniel Pakkala. Reference architecture and classification of technologies, products and services for big data systems. *International Journal of Information Management 37, 750–760*, 2017. DOI: 10.1016/j.bdr.2015.01.001.

[21] Qinglin Qi and Fei Tao. Digital twin and big data towards smart manufacturing and industry 4.0: 360 degree comparison. *IEEE Access*, 6:3585–3593, 2018.

[22] Paul Ralph, Nauman bin Ali, Sebastian Baltes, Domenico Bianculli, Jessica Diaz, Yvonne Dittrich, Neil Ernst, Michael Felderer, Robert Feldt, Antonio Filieri, et al. Empirical standards for software engineering research. *arXiv preprint arXiv:2010.03525*, 2020.

[23] Amr T. Sufian, Badr M. Abdullah, Muhammad Ateeq, Roderick Wah, and David Clements. A roadmap towards the smart factory. In *2019 12th International Conference on Developments in eSystems Engineering (DeSE)*, pages 978–983, 2019.

[24] Khalid Hasan Tantawi, Ismail Fidan, and Anwar Tantawy. Status of smart manufacturing in the united states. In *2019 IEEE 9th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0281–0283, 2019.

[25] Fei Tao, Qinglin Qi, Ang Liu, and Andrew Kusiak. Data-driven smart manufacturing. *The Society of Manufacturing Engineers, 0278-6125*, 2018. DOI: 10.1016/j.jmsy.2018.01.006.

[26] Birgit Vogel-Heuser, Thomas Bauernhansl, and Michael Ten Hompel. Handbuch Industrie 4.0 Bd. 4. *Allgemeine Grundlagen*, 2, 2020.

[27] Jonathan Stuart Ward and Adam Barker. Undefined by data: a survey of big data definitions. *arXiv preprint arXiv:1309.5821*, 2013.

[28] Stefan Windmann, Alexander Maier, Oliver Niggemann, Christian Frey, Ansgar Bernardi, Ying Gu, Holger Pfrommer, Thilo Steckel, Michael Krüger, and Robert Kraus. Big data analysis of manufacturing processes. *12th European Workshop on Advanced Control and Diagnosis (ACD 2015), Journal of Physics: Conference Series 659 (2015)*, 2015. DOI: 10.1088/1742-6596/659/1/012055.

[29] Wolfram Wingerath, Felix Gessert, Steffen Friedrich, and Norbert Ritter. Real-time stream processing for big data. *it - Information Technology, 58(4), 186-194.*, 2016. DOI: 10.1515/itit-2016-0002.

[30] Pai Zheng, Zhiqian Sang, Ray Y Zhong, Yongkui Liu, Chao Liu, Khamdi Mubarok, Shiqiang Yu, Xun Xu, et al. Smart manufacturing systems for industry 4.0: Conceptual framework, scenarios, and future perspectives. *Frontiers of Mechanical Engineering*, 13(2):137–150, 2018.

[31] Ray Y Zhong, Stephen T Newman, George Q Huang, and Shulin Lan. Big data for supply chain management in the service and manufacturing sectors: Challenges, opportunities, and future perspectives. *Computers & Industrial Engineering*, 101:572–591, 2016.