Proceedings of the Federated Conference on Computer Science and Information Systems pp. 427–431
DOI: 10.15439/2019F310
ISSN 2300-5963 ACSIS, Vol. 18

# Towards Big Data Solutions for Industrial Tomography Data Processing

Aleksandra Kowalska[1], Piotr Łuczak[2], Dawid Sielski[3], Tomasz Kowalski[4],
Andrzej Romanowski[5] and Dominik Sankowski[6]
Institute of Applied Computer Science, *Lodz University of Technology*
Łódź, Poland
[1]akowalska@kis.p.lodz.pl, [2]pluczak@kis.p.lodz.pl, [3]dawid.sielski@outlook.com, [4]t.kowalski@kis.p.lodz.pl,
[5]androm@kis.p.lodz.pl, [6]dsan@kis.p.lodz.pl

***Abstract*—This paper presents an overview of what Big Data can bring to the modern industry. Through following the history of contemporary Big Data frameworks the authors observe that the tools available have reached sufficient maturity so as to be usable in an industrial setting. The authors propose the concept of a system for collecting, organising, processing and analysing experimental data obtained from measurements with process tomography. Process tomography is used for noninvasive flow monitoring and data acquisition. The measurement data is collected, stored and processed to identify process regimes and process threats. Further general examples of solutions that aim to take advantage of the existence of such tools are presented as proof of viability of such approach. As the first step in the process of creating the proposed system, a scalable, distributed, containerisation-based cluster has been constructed, with consumer-grade hardware.**

***Index Terms*—Big Data, Process Tomography, data processing, data acquisition**

## I. Introduction

MEASUREMENT technologies are a practical challenge for engineers who are increasingly improving them, with novel algorithms for solving optimisation problems being continuously developed.

With the advent of industry 4.0 [1], the volume of the measurement data generated by industrial process becomes too large for processing on a single workstation. Through the use of wireless sensor networks, measurements can be taken in places previously inaccessible for traditional, wired solutions, hence allowing for preventative repair of equipment before the actual failure occurs [2]. As such new, sensor-rich systems are created there exists an unprecedented opportunity to derive deeper insights regarding the nature of the whole process from the large volume of collected data.

An example of a system rich in sensors is process tomography. It is a rapidly developing non-invasive diagnostic technique, finding wider and wider applications in various fields of science. In recent years, process tomography applications have been developed in the petrochemical, pneumatic and gravitational transport of bulk materials, as well as in the pharmaceutical industry and biomedicine. Fig. 1 shows different types of process tomography systems.

In the issues of non-invasive diagnostics and monitoring, various types of tomographic sensors can be combined to provide multi-modality, versatility and adaptation to the dynamics of the industrial process. The tomographic systems during computer diagnostics of the industrial process can also be supported by additional measurement sources, including ultra-fast video cameras, dedicated flow meters, scales, pressure and temperature sensors. In this way, additional information about the flow and its parameters is obtained. In addition, measurements from tomographic systems can be further used to reconstruct two- and three-dimensional images - both raster sequences and movie sequences. If this data comes from many sensors at the same time, is collected with high time resolution, and the industrial process is long-lasting, a large amount of data appears, which needs to be structured and categorised in appropriate database structures.

The diagnostic information collected using the process tomography measurement systems over a longer period of time can be characterised by the size of even a dozen terabytes, which predestines it for the term Big Data. In particular, such a term can be defined as long-term acquisitions of three-dimensional tomographic images with high temporal-spatial resolution, video recordings, sets of mathematical-physical models and descriptions of experiments. Therefore, this paper presents the concept of a system for collecting, organising, processing and analysing experimental data obtained from measurements using process tomography. The results of conducted research will enable computer systems of non-invasive industrial diagnostics to make quick and reliable decisions in the field of monitoring and control.

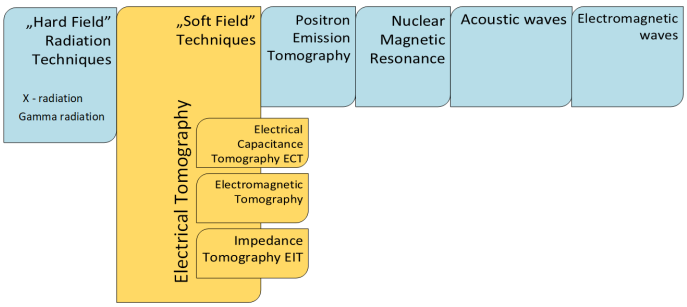


Fig. 1. Types of process tomography measurement modalities. Second from the left illustrates electrical tomography systems.

The contribution of this paper is the proposal for an ar-

chitecture based on containerised instances of Apache Spark and Hadoop which promises to offer soft real-time processing capabilities, whilst being trivially extendable to any number of distributed machines. Since all services in the proposed system are containerised, it is capable of being expanded or shrunk to fit the current needs through the simple process of either starting or shutting down a number of container instances.

## II. Big Data in industrial applications

### A. Evolution of Big Data solutions

The Big Data technology as known today, begun with the creation of the Google File System which enabled large scale data storage while using low-cost commodity hardware [3]. While providing no distributed computational capability, it contributed fault tolerant, multi-access file storage which facilitated concurrent work on the same file, hence implementing distributed producer-consumer queues for large (Multi-GB) files. This architecture, albeit limited, provided the foundation for the systems capable of handling both high volume data at rest (static files) and in motion (dynamically appended files), thus inherently satisfying both the Volume and Velocity part of the five V's of Big Data [4].

The next step in the development of a full Big Data stack was the MapReduce paradigm [5]. This paradigm was created as a means of unifying the disparate solutions that previously had to split their focus between the problems of pluralisation and the actual problem hence obscuring the latter. This new paradigm was a reaction to this unwanted, additional complexity and provided a simple abstraction layer allowing the programmers to specify their computations in the form of a chain of *map* and *reduce* steps, hence its name. The *map* function is responsible for processing the input key-value pair into intermediate results, also in the form of a key-value pair, whilst the *reduce* step is responsible for merging the intermediate values into the final result. This system, while capable of operating separately from the distributed file system, generally exists in symbiosis with the storage nodes of such a file system, thus resulting in the popular technique used to remove bottlenecks in the data processing, which generally referred to as "bringing processing to the data". The combined GFS and MapReduce provided the foundation for modern Big Data processing systems capable of satisfying all five V's, namely the aforementioned Volume and Velocity, and the previously unmentioned Veracity, Variety and Value.

Whilst the initial advancements in distributed file storage were proprietary and generally only available in the companies on the bleeding edge of Big Data technology, the Hadoop Distributed File System alongside with the rest of the Hadoop project were created by the Apache foundation, thus resulting in a significantly lowered barriers to entry into the Big Data industry [6]. The Open Source nature of this project also provided a good basis for the development of new data processing tools hence becoming an unofficial standard for big data storage and analysis systems.

One of the limitation of the early MapReduce systems was their inherent assumption that the data flow is acyclic, that is that a computational task begins with data being read from the file system and ends with the resulting data being written back to said file system. This, combined with the fact that distributed file systems, such as the GFS and HDFS were optimised for high throughput and not low file access latency, meant that, while possible, the execution of tasks that reused the working set of data was inadvertently slowed down by the file systems access time. In order to facilitate the execution of such tasks, a new framework named Spark was proposed [7]. Spark's Resilient Distributed Datasets (RDDs) provided a way for the data to be cached in memory during the execution of the iterative task, hence avoiding the performance penalty resulting from repeated file system accesses. This novel approach allowed for a significant decrease in the execution time of iterative tasks such as linear regression or the alternating least squares computation, making Big Data based machine learning solutions viable.

With the proliferation of new data processing frameworks the tight coupling between the MapReduce programming model and the Hadoop file system became an impediment which had to be worked around by the developers of these solutions. A new incarnation of the Hadoop framework was created with the aim of alleviating the aforementioned limitations and clearly separating the resource management from the programming model. The framework was named Yet Another Resource Negotiator (YARN) [8]. In this framework, MapReduce was no longer the primary focus and thus became simply one of the possible tools that could be run. This new version of MapReduce is commonly referred to as MapReduceV2. A multitude of programming frameworks, that initially had to be built on top of MapReduce and thus were beholden to its limitations, were updated to use this new model even before YARN left the beta stage of its development, therefore validating the design decisions made by the YARN creators.

As a result of the aforementioned progress the Big Data ecosystem became not only a feature rich set of tools but also a stable and mature foundation for building new solutions and applications that deal with large volumes of data [9].

### B. Industrial applications

Modern industry tends to generate an enormous amount of data every day while in many cases lacking the technical capabilities required effectively process and derive long-term insights from it [10] [11]. In many cases the extent of use of this data boils down to being shown to the operator on-site so as to facilitate the monitoring of the industrial process. Such analysis is not only limited by the finite capabilities of both human mind and memory but also results in very low bus factor, making the expert an unexpendable component of the process. This situation however is slowly changing with new approaches and initiatives such as the industry 4.0.

The financial industry has devoted considerable research efforts into what could be considered proto-Big Data processing since as early as 1997 [12]. In this sector one of the most important computational tasks is fraud detection, which unlike most statistical data analysis, concerns itself with

data in motion, with real time emphasis. Some AI application require enormous collections of data to be stored, prepared and processed by machine learning such as the EEG annotation data [13], industrial emergency states detection [14] [15] or medical augmented reality future diagnosis [16].

## III. State of technology

As presented in II-B there exist numerous applications of big data in different domains. One of the interesting examples was reported by Skuza et al. [17], another example of the barely explored domain is the usage of big data for process tomography [18]. Process tomography is a set of techniques that are responsible for acquiring data from a given process, processing them and providing information about concentration distribution and flow.

Process tomography together with big data solves lots of different problems that emerge with traditional approach. Romanowski et al. started to join these two fields to propose a solution for detecting material plugs in pneumatic conveying measurement data [18] [19]. The authors used Hadoop (installed on three independent computers) together with Mahout machine learning library. The data consisted of five experiments performed on several different pipe diameters with different flow conditions. The tests were performed in two distinct locations at different times. Total data obtained during the data acquisition part was about 40 GB but it must be stated that the computational system is able to handle significantly larger amount of data. The authors present proof that the combination of big data tools and conventional algorithms can be used in automatic detection of material plugs in vertical or horizontal flows. This indicates that using big data tools such as Hadoop is an excellent alternative for traditional approach.

Process tomography uses lots of different techniques which include lots of different sensors from which the data is obtained. Numerous different sensors with numerous output format are a problem designed to be solved by big data. Rymarczyk et al. propose and discuss the system which is designed to optimize and automate given process by using numerous tomographic sensors [20]. The paper indicates the importance of using such systems for maintaining competitiveness. The design of the system includes multiple sensors like ECT (Electrical Capacitance Tomography), ERT (Electrical Resistivity Tomography), UST (UltraSound Tomography), together with temperature and pressure information. In the next step the authors perform the data acquisition during which the data is saved to a server for later processing. Next part uses image reconstruction together with cloud computing which later is employed to control industrial processes.

There are numerous steps needed to be performed in the system in order to obtain data required to control some process. There is also an increase in amount of data that needs to be processed. As a result the data processing part may take a longer amount of time. To lower that the parallelisation can be introduced. This approach was used by Chen et al. where he designed and tested solution that uses crowd-sourcing to help with understanding the particle-tracking problem [21]. The authors addressed that at the time being there are no solutions that can accurately analyse flow in silos. This resulted in building a system that employed experts and non-experts to analyse tomographic images. The results from this system were compared to the results from the automatic approach. The paper clearly states that the crowd solution is significantly better in terms of scale, delivered result and economics in comparison to the automatic approach. What is more, the system can be applied in different domains which makes it generalisable.

The goal is to produce system that will be responsible for controlling industrial process in order not only to prevent undesirable behaviour from happening but also to obtain better understanding about the process along with the possibility of optimisation of the process. The newest approach proposed by Romanowski gives more insight on joining big data with process tomography by extending work presented by Chen and is an example of the desired system [4]. The author in his work designed a methodology that improves current state of knowledge, regarding recognition of specific bulk flow regimes (e.g. pipeline blockage threats). To obtain such improvement the author presented solution using Hadoop platform for distributed data processing on large quantities of data using cluster computing. For the classification part the Apache Spark with SVM as an algorithm was used as a supervised learning method. The system can also be employed to handle processes from different domains. The main advantage of this solution is the improvement in finding flow regimes that may be critical for the process and perform actions to prevent such flow. This work also describes how to jointly analyse incoherent data.

## IV. Goals, vision and proposed approach

The aim of the work is to develop new algorithms for the collection, organisation, processing and analysis of large data sets obtained from industrial systems of non-invasive diagnostics and control, based on process tomography.

As part of the work, it is proposed to use a containerisation-based [22] implementation system that will allow for easy increase of available disk space and computing power, both using an external cloud infrastructure and generally available consumer computer equipment. Due to the particular requirements of the platform components, it may prove inadvisable for such an approach to cover the entire project. The use of containerisation has certain consequences and limitations that should be considered. Containers introduce another (though lightweight) layer "reflecting" a given element of the platform from the concept of the real-time system. Stations and diversified installation software have specific hardware requirements (eg. control interfaces) or visualisation (eg. the use of calculations directly from the GPU), which may potentially disqualify the software as a service subject to containerisation. In connection with the above, work on the implementation system based on the above approach has been focused on the computational cluster responsible for Big data analysis of historical data and eventually, real-time sensor readings. Fig. 2 presents the proposed model in which the

dependence of the control loop on the cloud is limited to "fine tuning" so as to avoid affecting the delay of regulation.

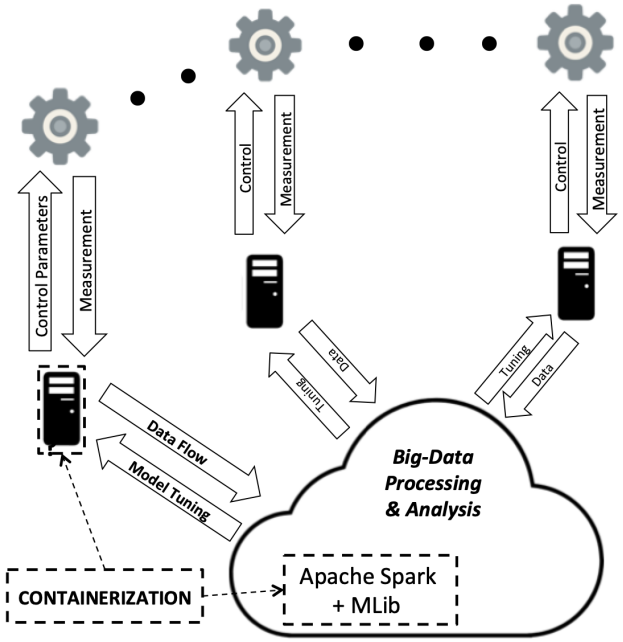


Fig. 2. Proposed model

Based on the aforementioned attempts, from section III, at creating data based systems for the purposes of analysis and control of tomographic processes, the planned focus of our research is twofold. Firstly, we plan to assess the benefit, defined as an increase in the computation speed of visualisation of the measured flow, of using a cluster consisting of consumer-grade hardware over the traditional approach of carrying out the computations on one high-end workstation. Secondly, we plan to assess the viability of using such cluster as a perpetually self-improving industrial controller. It is our belief that, provided enough training data, such a system could potentially operate with only minimal supervision from an operator, hence allowing a single expert to oversee a larger number of apparatus than it is currently possible. This concept will require an extensive amount of experimentation before it can be considered ready for industrial use, though at this early stage two criteria of viability can be defined:

1) capability of gradual auto-tuning of traditional controller
2) nearly real-time performance of the flow analysis

Should only one of these criteria be met the system would still be a valuable enhancement of a traditional system, either as an "auto-tuner" or as a real-time (or almost real-time), flow visualisation platform.

Should both of these criteria be met the future steps would include a "run-ahead", real-time simulation that could be used to predict potential faults and act accordingly to prevent them. An alternative approach for this purpose is also considered, namely instead of simulating the flow, an additional machine learning model could be used as a fault predictor. The architecture of the proposed cluster can be seen on figure 3.

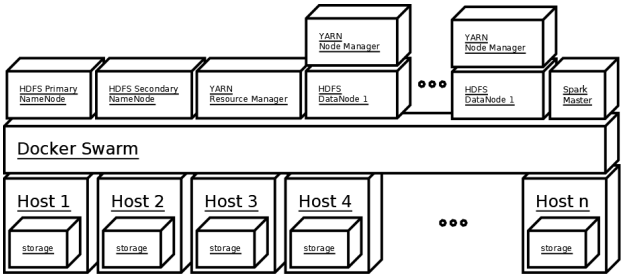


Fig. 3. Structure of the proposed cluster

## V. Evaluation

The initial evaluation of the system generated promising results, as the even with as few as 5 low cost machines it was capable of outperforming the existing single-machine solution present in the Tom Dyakowski Process Tomography Laboratory in Lodz University of Technology, shown on figure 4. In the initial testing, which used a very limited number of machines with the exact same hardware, the increase in the speed of computations (measured as a decrease in computation time) was linear, whilst the latency did not exceed 2.3 seconds. Due to a very limited number of machines with the exact same specifications, subsequent tests had to involve a variety of configurations. Since the machines available did not have the

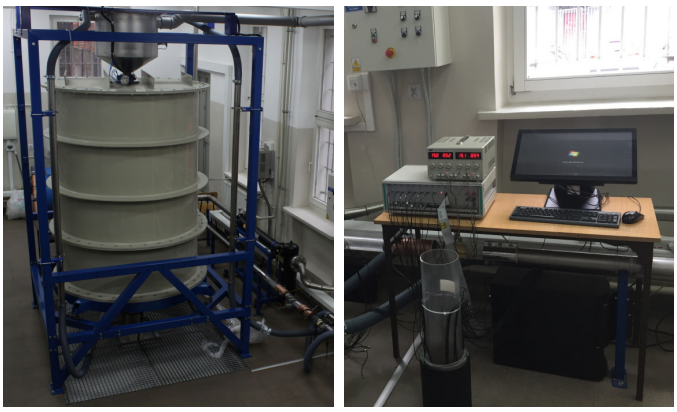

Fig. 4. Tom Dyakowski Process Tomography Laboratory: experimental silo (left) and data acquisition, processing and visualisation workstation (right)

exact same hardware specification, it is impossible to measure a precise number of machines at which the cluster becomes a more efficient computational environment, however this non-uniformity proves to be a notable advantage of our approach. Since the cluster was capable of working with machines with differing specifications, one could construct such a solution without incurring additional cost, by simply using the spare, possibly outdated, hardware already available on-site.

## VI. Discussion and future work

This paper presents the idea of a big data system based on tomographic solutions. The general structure of the cluster is shown in Figure 3. Particularly noteworthy here is the ease of expanding the system described, both in terms of

physical devices and additional services. An example of such an extension is the planned addition of more efficient nodes equipped with an increased amount of volatile memory and a high-end graphics card in order to facilitate parallel computing.

As part of further work, proprietary algorithms for automatic acquisition, analysis and interpretation of large data sets [23] will be developed, as well as intelligent decision algorithms supporting effective diagnostics and monitoring of flow processes [14]; including those to be the next step after crowdsourcing data labelling cases impossible to be automatically processed at once [24] [25]. The prepared algorithms will be verified both by simulation and by real experimental data from their own experiments carried out on the basis of unique, semi-industrial research installations. The results of research will enable computer systems to make quick decisions in the field of monitoring and control. Furthermore, the computational environment may be suited to personal-medical applications, such as the continuous glucose monitoring, resulting in better diagnosis, safety and hence better life comfort [26].

## VII. Conclusion

This work presents a pioneering concept for the Big Data system. The solution concerns the integration of data from sensors used in process tomography, which enables imaging and feedback associated with control in the field of process optimisation. The presented infrastructure can bring tangible benefits in various sectors of the industry due to its scalable nature, allowing for smooth expansion as the company or its requirements grow. The described approach is based on a Docker Swarm cluster which facilitates easy fail-over in case of hardware node failures.

## Acknowledgment

This work is partially financed by the Smart Growth Operational Programme 2014-2020 project no POIR.04.01.02-00-0089/17-00. The project is conducted in the Institute of Applied Computer Science at the Lodz University of Technology.

## References

[1] H. Lasi, P. Fettke, H.-G. Kemper, T. Feld, and M. Hoffmann, "Industry 4.0," *Business & Information Systems Engineering*, vol. 6, no. 4, pp. 239–242, 2014. doi: 10.1007/s12599-014-0334-4

[2] V. C. Gungor, G. P. Hancke *et al.*, "Industrial wireless sensor networks: Challenges, design principles, and technical approaches." *IEEE Trans. Industrial Electronics*, vol. 56, no. 10, pp. 4258–4265, 2009. doi: 10.1109/TIE.2009.2015754

[3] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The Google file system," p. 29, 2003.

[4] A. Romanowski, "Big data-driven contextual processing methods for electrical capacitance tomography," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1609–1618, 2019. doi: 10.1109/TII.2018.2855200

[5] J. Dean and S. Ghemawat, "Mapreduce: simplified data processing on large clusters," *Communications of the ACM*, vol. 51, no. 1, pp. 107–113, 2008. doi: 10.1145/1327452.1327492

[6] K. Shvachko, H. Kuang, S. Radia, and R. Chansler, "The hadoop distributed file system," in *Mass storage systems and technologies*, 2010. doi: 10.1109/MSST.2010.5496972 pp. 1–10.

[7] M. Zaharia, M. Chowdhury, M. J. Franklin, S. Shenker, and I. Stoica, "Spark: Cluster computing with working sets." *HotCloud*, vol. 10, no. 10-10, p. 95, 2010.

[8] V. K. Vavilapalli, A. C. Murthy, C. Douglas, S. Agarwal, M. Konar, R. Evans, T. Graves, J. Lowe, H. Shah, S. Seth *et al.*, "Apache hadoop yarn: Yet another resource negotiator," in *Proceedings of the 4th annual Symposium on Cloud Computing*, 2013. doi: 10.1145/2523616.2523633

[9] P. Basanta-Val, N. C. Audsley, A. J. Wellings, I. Gray, and N. Fernández-García, "Architecting time-critical big-data systems," *IEEE Transactions on Big Data*, vol. 2, no. 4, pp. 310–324, 2016. doi: 10.1109/TBDATA.2016.2622719

[10] K. Grudzien, A. Romanowski, D. Sankowski, and R. A. Williams, "Gravitational granular flow dynamics study based on tomographic data processing," *Particulate Science and Technology*, vol. 26, no. 1, pp. 67–82, 2007. doi: 10.1080/02726350701759373

[11] T. Rymarczyk, "Using electrical impedance tomography to monitoring flood banks," *International Journal of Applied Electromagnetics and Mechanics*, vol. 45, pp. 489–494, 2014. doi: 10.3233/JAE-141868

[12] K. Grudzien, A. Romanowski, and R. A. Williams, "Application of a bayesian approach to the tomographic analysis of hopper flow," *Particle & Particle Systems Characterization*, vol. 22, no. 4, pp. 246–253, 2005. doi: 10.1002/ppsc.200500951

[13] S. Opałka, B. Stasiak, D. Szajerman, and A. Wojciechowski, "Multi-channel convolutional neural networks architecture feeding for effective eeg mental tasks classification," *Sensors*, vol. 18, no. 10, 2018. doi: 10.3390/s18103451

[14] A. Romanowski, "Contextual processing of electrical capacitance tomography measurement data for temporal modeling of pneumatic conveying process," in *Proceedings of the 2018 Federated Conference on Computer Science and Information Systems*, vol. 15, 2018. doi: 10.15439/2018F171 pp. 283–286.

[15] K. Grudzien, A. Romanowski, and R. A. Williams, "Application of a bayesian approach to the tomographic analysis of hopper flow," *Particle & Particle Systems Characterization*, vol. 22, no. 4, pp. 246–253, 2005. doi: 10.1002/ppsc.200500951

[16] A. Nowak, M. Wozniak, M. Pieprzowski, and A. Romanowski, "Towards amblyopia therapy using mixed reality technology," in *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2018. doi: 10.15439/2018F335 pp. 279–282.

[17] M. Skuza and A. Romanowski, "Sentiment analysis of twitter data within big data distributed environment for stock prediction," in *2015 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 2015. doi: 10.15439/2015F230 pp. 1349–1354.

[18] A. Romanowski, M. Skuza, P. Wozniak, K. Grudzien, and Z. Chaniecki, "Big data computational environment for tomography measurement data," *Process Tomography WCIPT7, Poland*, 2013.

[19] A. Romanowski, K. Grudzien, Z. Chaniecki, and P. Woźniak, "Contextual processing of ECT measurement information towards detection of process emergency states," in *13th International Conference on Hybrid Intelligent Systems*, 2013. doi: 10.1109/HIS.2013.6920448 pp. 291–297.

[20] T. Rymarczyk and J. Sikora, "Applying industrial tomography to control and optimization flow systems," *Open Physics*, vol. 16, p. 46, 2018. doi: 10.1515/phys-2018-0046

[21] C. Chen, P. W. Woźniak, A. Romanowski, M. Obaid, T. Jaworski, J. Kucharski, K. Grudzień, S. Zhao, and M. Fjeld, "Using crowdsourcing for scientific analysis of industrial tomographic images," *ACM Trans. Intell. Syst. Technol.*, vol. 7, no. 4, p. 52, 2016. doi: 10.1145/2897370

[22] W. Felter, A. Ferreira, R. Rajamony, and J. Rubio, "An updated performance comparison of virtual machines and linux containers," in *2015 IEEE International Symposium on Performance Analysis of Systems and Software*, 2015. doi: 10.1109/ISPASS.2015.7095802 pp. 171–172.

[23] A. Kowalska, R. Banasiak, A. Romanowski, and D. Sankowski, "3d-printed multilayer sensor structure for electrical capacitance tomography," *Sensors*, vol. 19, no. 15, 2019. doi: 10.3390/s19153416

[24] A. Romanowski, P. Łuczak, and K. Grudzień, "X-ray imaging analysis of silo flow parameters based on trace particles using targeted crowdsourcing," *Sensors*, vol. 19, no. 15, 2019. doi: 10.3390/s19153317

[25] I. Jelliti, A. Romanowski, and K. Grudzień, "Design of crowdsourcing system for analysis of gravitational flow using x-ray visualization," in *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems*, M. Ganzha, L. Maciaszek, and M. Paprzycki, Eds., vol. 8, 2016. doi: 10.15439/2016F288 pp. 1613–1619.

[26] P. Kucharski, K. Pagacz, A. Szadkowska, W. Młynarski, A. Romanowski, and W. Fendler, "Resistance to data loss of glycemic variability measurements in long-term continuous glucose monitoring," *Diabetes Technology & Therapeutics*, vol. 20, no. 12, pp. 833–842, 2018. doi: 10.1089/dia.2018.0247