A systemic mapping of methods and tools for performance analysis of data streaming with containerized microservices architecture

Simone Ris*, Jean Araujo*†, and David Beserra‡
*Departamento de Computação, Universidade Federal de Sergipe, São Cristóvão, Brazil
†Universidade Federal do Agreste de Pernambuco, Garanhuns, Brazil
‡École Pour l'Informatique et les Techniques Avancées (EPITA), Paris, France
simonerss@dcomp.ufs.br*, jean.teixeira@ufape.edu.br†, david.beserra@epita.fr‡

***Abstract*—With the Internet of Things (IoT) growth and customer expectations, the importance of data streaming and streaming processing has increased. Data Streaming refers to the concept where data is processed and transmitted continuously and in real-time without necessarily being stored in a physical location. Personal health monitors and home security systems are examples of data streaming sources. This paper presents a systematic mapping study of the performance analysis of Data Streaming systems in the context of Containerization and Microservices. The research aimed to identify the main methods, tools, and techniques used in the last five years for the execution of this type of study. The results show that there are still few performance evaluation studies for this system niche, and there are gaps that must be filled, such as the lack of analytical modeling and the disregard for communication protocols' influence.**

***Index Terms*—data streaming, container, microservice, reali-bility, performance, availability**

I. Introduction

Thanks to the continuous evolution of network technologies, an increasing number of devices are interconnected every day. As a result, there is a need to develop ways to manage the enormous amount of data generated by these devices on a daily basis [1]. Some solutions in this context involve processing and transmitting data continuously and in real-time, without necessarily storing it permanently. This approach is known as data streaming [2]. In such systems, data is modeled as transient data streams [3]. Applications of data streaming can be found in various fields, including finance, network monitoring, security, telecommunications data management, manufacturing monitoring, and sensor networks.

The performance of data streaming systems is critical since these systems deal with a vast volume of data and require fast and accurate processing [4]. As computing structures become more distributed, they can grow considerably in size to encompass the cloud. Details such as the size of the input flow, the capacity to receive data, or even the processing structures need to be architected effectively, or the computing capacity of the entire system can be compromised, resulting in a loss of accuracy [2].

Dantas et al. [5] have highlighted that the lack of accuracy in cloud-based services - constant lack of service, failures in reliability and security - leads to negative user experiences, resulting in losses to the service provider. Thus, service providers need to ensure high levels of availability and reliability to maintain user satisfaction. One way to accomplish this is by using models to represent the performance of systems, which enables reliable results at a very low cost and eliminates the need to build a real system to evaluate it [6].

This paper aims to report research on the main materials, methods, and techniques used to perform performance evaluations of data streaming systems. This will help us understand the methodology already applied and establish what can be added to plan these services more effectively. We have chosen the research modality called Systematic Mapping (SM) for this state-of-the-art study. This type of study seeks to synthesize all available knowledge on a research topic [7]. It involves categorizing and locating studies related to the topic and creating a knowledge base about it. In this process, one can even identify the lack of information on the topic, indicating the need for more primary studies related to the topic.

This paper is organized as follows: Section II provides an overview of data streaming systems, including their main concepts and the importance of efficient planning. Section III presents the research protocol and the steps followed in this study, along with the results obtained at each stage. Section IV analyzes the extracted data and identifies the trends identified in the mapping. Finally, Section V provides an overview of the acquired knowledge and outlines the next steps for modeling data streaming systems more efficiently.

II. Data Streaming

The previous section introduced the importance of managing data generated by interconnected devices and how data streaming is used to process and transmit data continuously and in real-time without necessarily storing it permanently. In this section, we provide further context by discussing the origins of data streaming and its unique characteristics.

The term ”data stream” was first introduced in 1998 in the paper ”Computing on Data Stream” [8], which proposed data streaming as a new type of data object. Today, data streaming has become a category of systems used in various contexts where a large amount of data generated needs to be processed quickly [4]. Unlike the traditional batch processing model, data streams arrive in different streams and from various sources through the internet, making it challenging to control the order in which data is processed [3].

Various factors must be considered in the design and planning of data streaming systems to ensure their accurate and reliable operation. For instance, the system needs to be able to handle different input sizes and receive data without any interruptions, while the processing structures must be designed to prevent any failures that could compromise the system's capacity [2]. Figure 1 provides a visual representation of how data streaming systems operate.

In summary, planning and designing a suitable architecture are crucial to ensuring the continuous and stable operation of data streaming systems. The next section will describe the research methodology used in this paper to identify the main materials, methods, and techniques employed to evaluate the performance of these systems.

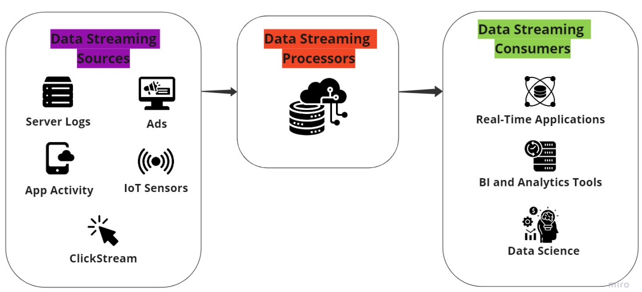

Fig. 1. Data Streaming Operation

III. Protocol Defined for This Systematic Mapping

This section reports the details of the steps taken in the execution of this mapping and the results obtained from each of these steps.

A. Research Questions

Considering the main interest of this systematic mapping, which was described in Section I, the main research question was elaborated as follows:

RS1: What are the best techniques that can be used to establish a highly efficient and effective model of Data Streaming system applied to Microservices and Containerization?

Factoring in the main aspects that may be compiled in this larger research question, one can arrive at the following secondary questions:

RS1.1: Which relevant data streaming applications have been proposed in the literature?

RS1.2: Which countries publish the most on data streaming?

RS1.3: What literature has proposed analytical modelling techniques?

RS1.4: What software tools have been used for prototyping?

RS1.5: What hardware tools have been used for prototyping?

B. Search Strategy

With the research questions established, it was possible to define that the search criteria capable of returning the studies necessary to obtain the intended state-of-the-art would be as follows:

1) Keywords: data streaming, data stream, container, microservice, Kubernetes, Docker, realibility, performance, availability

2) Search String: (reliability OR availability OR performance) AND (microservice OR container OR Kubernets OR Docker) AND ("data streaming" OR "data stream")

3) Databases of Research Used: As it is intended to bring together the relevant primary studies and related technologies that can be used, it was decided that it is suitable for this work to gather the patent registries of technologies associated with the scope of this research. Table I lists the search databases used in this research.

TABLE I
List of Research Bases

Research Base	URL
Scopus	https://www.scopus.com
MDPI	https://www.mdpi.com
IEEE Xplore	https://ieeexplore.ieee.org
Web of Science	https://www.webofscience.com
ACM Digital Library	https://dl.acm.org/

C. Search Results of Relevant Works

After performing the searches in the selected databases, it was possible to identify a total of 91 primary studies.

TABLE II
Search Results in Scientific Research Bases

Research Base	Found	Duplicate
Scopus	35	4
MDPI	22	0
IEEE Xplore	15	3
Web of Science	14	5
ACM Digital Library	5	0

As shown in Table II, at first identified a total of 12 duplicate works. Thus, 79 of these publications proceed to the selection stage, described in Section III-D.

D. Criteria for the Selection of Publications

The guidelines of [7] specify that the selection step is a screening process where irrelevant publications are removed from the set of primary studies obtained in the search step. This screening is performed by applying the selection criteria decided during the definition of the protocol in the studies [9]. Thus it is expected that only the works that provide direct evidence on the research issue are used for data extraction. The criteria that were defined for the protocol of this research were the following:

Inclusion Criteria:

1) The publication shall describe performance evaluations of data streaming systems;
2) The publication shall describe methods, techniques or protocols for data streaming systems;
3) The purpose of the paper should be in the context of describing hardware or software tools to establish data streaming services.

Exclusion Criteria:

1) The publication does not address methods, techniques or performance evaluations for the improvement of data streaming systems;
2) The publication does not present full texts of the studies;
3) The publication are unavailable for download on the Internet;
4) The publication is not in English.

E. Results of the Selected of Papers

Following the guidelines of [9], the publications were selected after eliminating duplicates in the search phase. The inclusion or exclusion of papers was decided based on reading their Abstract, Introduction, and/or Conclusion. In total, 33 studies were included, and the proportion of studies included by research basis is shown in Figure 2.

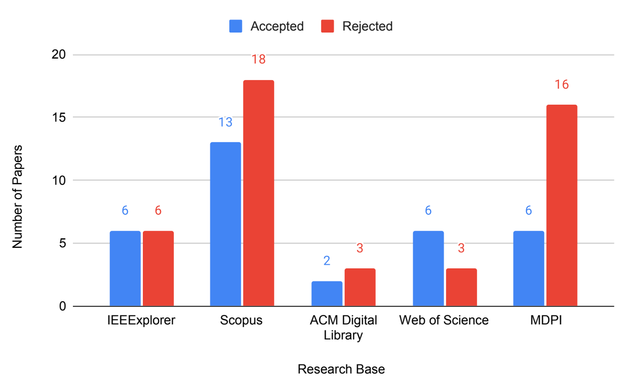

Fig. 2. Results of the Selection Stage

F. Quality Assessment

In addition to the general inclusion/exclusion criteria, evaluating the *Quality Assessment Checklist* can guide the interpretation of the findings and determine the importance of this for this study and future research [9]. The following quality criteria have been defined, which relate to the fundamental subject of mapping:

1) The main study of the publication should be focused on data streaming systems with microservices architecture;
2) The main study of the publication should be focused on data streaming systems that apply containerization.

Publications with the status of 'Ranked' in the checklist item are scored 1. Those who fall in the status of 'Partially Classified' receive score 0.5, and those who are considered 'Not Classified' in the item, receive score 0. The cut grade for this step is score 1. This step was performed by applying the technique of scanning and skimming, where the text is observed in general, in a reading that does not delve into the details, just searching for keywords.

The quality check results are expressed proportionally in Figure 3. 21.2% of publications that studied data streaming systems but did not address applications built with microservices architecture and did not deal with containerization were deleted, reaching a score of 0. A total of 27.3% of the studies addressed the issue of containerization without mentioning the aspects of the architecture of microservices, and 6.1% of the papers dealt with microservices outside the context of containerization. These reached a 1.0 score. Finally, 45.5% of the 33 papers selected received enough scores to proceed to the selection stage, as they dealt specifically with data streaming systems in the context of containerization of applications with microservices architecture.

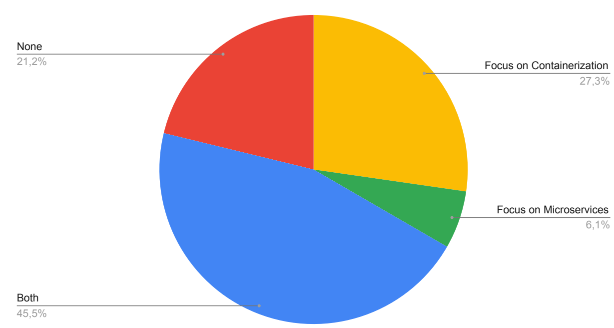

Fig. 3. Quality Assessment

G. Form of Extraction of Data

[9] defines that data extraction forms should be drawn up to gather the data necessary to answer the research questions identified for the study so that data extraction establishes a set of numerical values that will be important for the interpretation of the results. Given this rule, the following questions were drawn up for the data extraction form:

1) What is the country of publication?
2) Which institution was the research linked to?
3) What is the general subject of the study?
4) What software tools were used for prototyping in the reported study?
5) What hardware tools were used for prototyping in the reported study?
6) Which communication protocols are cited and/or tested in the publication?
7) What performance analysis metrics are cited and/or tested in the publication?
8) What analytical modelling techniques are cited in the publication?

Once planned the details of the execution of the mapping, it was followed for the execution of the steps, so that the obtained results will be described in the subsequent sections.

IV. Result Analysis

This section presents and summarizes the analysis of the papers we analysed on this systematic review.

A. Which countries publish about data streaming?

Among the texts selected for the complete reading and synthesis of knowledge, five were published with German institutions' participation. In second place came the United States of America and Italy, with three papers published each. The third place in the ranking of classified publications is occupied by Saudi Arabia, with two articles. Finally, there were works from China, United Arab Emirates, Spain, India, Iraq, Ireland and Pakistan, and each of these countries had one publication included in this research. The participation of each country in the selected studies is shown in the graph in Figure 4.

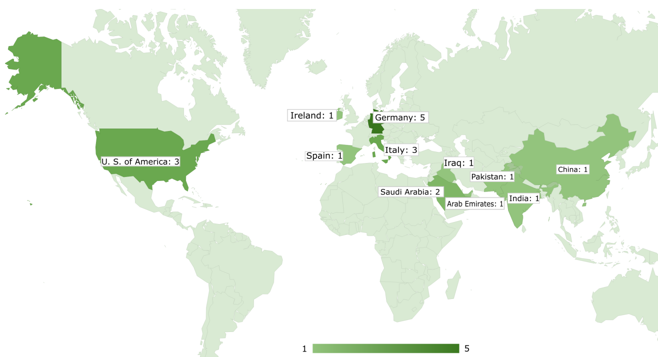

Fig. 4. Qualified Studies by Country

In the case of the institutions to which the work was linked, it should be noted that at least three research institutes directly related to the industry were present, in addition to a health institution: *IBM Research*, *LinkedIn Corp*, *NEC Laboratories America*, *Institute for Innovative Mikroelektronik* and *Tangshan Gongren Hospital*. This fact can be understood as an indication of the interest of the technology market in this type of system. An interesting example to be cited is the work of [10], where it was studied the improvement of wearable devices with IoT technology for monitoring patients with hyperthyroidism, showing the potential that improvements in data streaming technologies contribute to the health of the general population. The other institutions present in the research are universities and higher education institutes. The organizations responsible for the studies qualified for this research were the following: University of Illinois, LinkedIn Corp, University of Potsdam , Freie Universität Berlin, NEC Laboratories America, Princeton, Annamacharya Institute of Technology & Sciences, King Saud University, Al-Nahrain University, King Faisal University, Free University of Bozen-Bolzano, University of Málaga, Institut for Innovative Mikroelektronik, University of Technology Cottbus-Senftenberg, Tangshan Gongren Hospital, Aachen University, University of Jeddah, National University of Sciences and Technology (Paquistão), Technische Universität Braunschweig, Baylor University, United Arab Emirates University, IBM Research and University of Bologna.

Regarding the distribution of classified works, by year of publication, it is illustrated in Figure 5, where it can be noted that the year in which most of these works were published in 2019, when there were in total 5 papers within the niche of this research. There was a drop in the number of publications in 2021, and the number of publications grows again in 2022, which may be an indication that the subject in question tends to resume relevance.

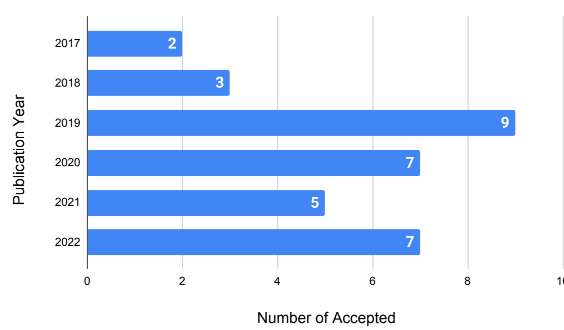

Fig. 5. Qualified Studies by Year of Publication

B. What software tools have been used for prototyping?

Some paper works were dedicated to proposing software to establish data streaming systems, they are: Apache Samza [11], DataXe [12], Kafka-ML [13], Internet-of-Things (IoT)-Infrastructure-as-a-Service (IoT-IaaS) [14] and TRAK [15]. The software cited in the publications, mostly used in the testbeds of the investigation,were listed in the Table III.

TABLE III
The softwares mostly used in the testbeds

Category	Software
Operating Systems (OS)	Ubuntu, SPIFFS, Hypriot, Kernel Linux and Firebase.
Virtual Machines	Docker Virtual Machine, OpenStack and Libvirt.
Platforms for Processing of Streaming	Apache Storm, Apache Kafka, Apache Beam, Apache Spark, Apache Spark Streaming, Apache Flink and Debezium.
Orchestrators of the Container	Apache Spark clusters, Trusted Orchestration Management (TOM) and Linux containers.
Platforms for Clusters	Docker, Docker Swarm, Apache Hadoop.
Brokers	Message Broker Redis.
Serves	Firebase and Zookeeper.
Utilities	Grafana, Open-JDK, Docker CLI, cAdvisor, Prometheus, Painel Horizon of OpenStack, Apache Hadoop YARN, Hadoop HDFS, MapReduce, ESPAsyncWebServer3, AssemblyScript4 and v4l2-loopback.

C. Communication Protocols Cited and/or Tested

At least four articles fail to cite protocols, remaining at the highest level of the application stack used for data streaming communication. The others only cite the protocols without going into details about the impact of the chosen protocol on the performance of the described solution. The list of protocols and the number of times they were cited in the 15 papers is shown in the graph in Figure 6.

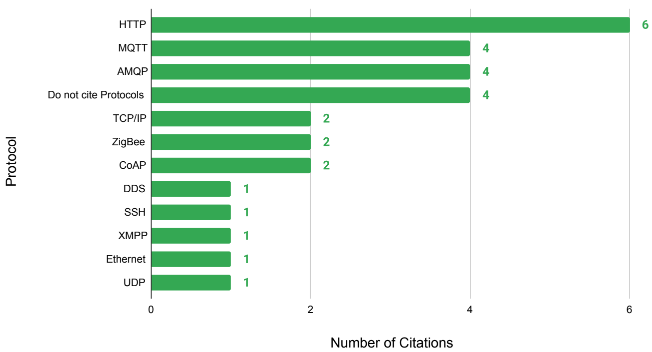

Fig. 6. Communication Protocols and Number of Citations Occurred

D. What hardware have been used for prototyping?

The Table IV list the hardware tools used in the test platforms of each analyzed study. Four works did not explain the hardware tools used, and they are: [16], [17], [18] and [13].

E. What analytical modeling techniques have been proposed by the literature?

It was found that no publications used analytical models to describe the performance of systems in their works. This fact represents a significant gap in the scope of data streaming systems performance research since these methods have significant advantages.

As explained by [25], a range of models can be used for the analytical evaluation of systems. The most cited examples are Petri nets, Reliability block diagrams, fault trees and Markov chains. They are mostly used to model fault tolerance in systems, availability and reliability evaluation, among others. These types of models offer high modeling power, so it has the potential to be an important tool in the task of establishing several performance characteristics in software systems treated in this work.

F. What are performance analytics metrics cited and/or tested in the publication?

The graph in Figure 7 shows the list of all metrics found in the papers studied and illustrates the number of times each was found in the works.

The analysis led to the finding that the metric *Processing Time* or *Runtime*, has been the greatest concern of researchers

TABLE IV
THE HARDWARE TOOLS USED IN THE TESTBEDS

Reference	Hardware Components
[11]	6 nodes Cluster YARN; 4 Resource Managers (RMs) and 2 Node Managers (NMs). Each NM was a cutting-edge machine with 64 GB of RAM, 24 CPUs core, 1 SSD of 1,6 TB, 2 HDDs of 1 TB and a network Ethernet full-duplex of 1 Gbps.
[19]	3 node Cluster with 64 GB main memory and CPU Intel(R) Xeon(R) E5-2697 v3 @ 2,60GHz with eight cores each. Machine with CPU Intel Core i9-7980XE, GPU NVIDIA GeForce RTX 2070 and 32 GB of main memory.
[12]	8 Raspberry pi2 model B, for the cluster. Config.: System on chip – BCM2836 Broadcom; Ethernet – 100Mb; Memory 1GB; CPU ARM Cortex-A7 quad-core 900MHz 32-bit.
[20]	8 Raspberry pi2 model B, to mount the cluster. Config.: System on chip – BCM2836 Broadcom; Ethernet – 100Mb; Memory – 1GB; CPU ARM Cortex-A7 quad-core 900 MHz 32-bit.
[10]	A microcontroller board ESP32; Modules of the TTGO and the DEVKIT V1.
[21]	Apple MacBook Pro; Onboard Computers: two Raspberry Pi2 and one Raspberry Pi3; Temperature-humidity sensor DHT11; Temperature sensor of the CPU; Webcam A4Tech.
[14]	Raspberry Pis; Desktop computers, laptops, servers; Smartphones or tablets; Traditional cloud service (in this implementation, Firebase).
[22]	8 nodes Raspberry Pi 2 – Model B (ARMv7; SoC Broadcom BCM2836; CPU 900 MHz quad-core 32-bit ARM Cortex-A7; RAM 1 GB; Ethernet 10/100Mbit/s). And an 8 GB microSD card for OS installation; Veracity Camswitch 8 Mobile switch; Power over Ethernet (POE); 10 port switch 10/100Mbit/s, with eight outputs 802.3 POE; category 5E SFTP cables.
[23]	PC equipped with CPU Intel(R) Core(TM) i7-8700 @ 3. 20GHz, 32 GB of RAM and 2 TB disks.
[24]	Development board Nvidia Tegra TK1; Board Raspberry Pi 2; 22 VMs each with 2 virtual CPUs and 2 GiBytes of DRAM.

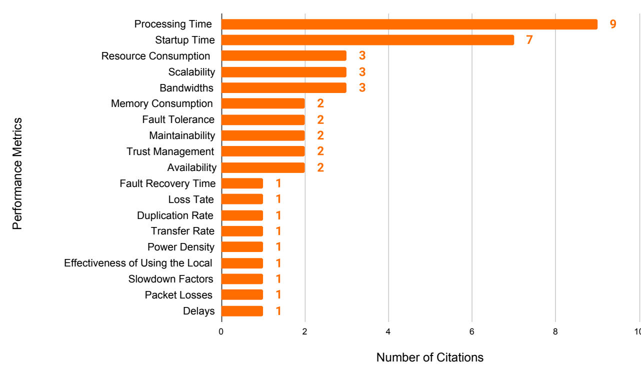

Fig. 7. Performance Metrics Found in Publications

in the performance evaluation of data streaming systems since it was the most cited mentioned at least nine times in the evaluated works. The second most cited metric was the *Startup Time*, which was worked seven times in the studies. *Resource*

Consumption, *Scalability*, and *Bandwidth* were mentioned three times each, being together in third place in the ranking of the most worked metrics in the studies selected for this research.

V. Conclusions

This paper has explored the performance analysis of Data Streaming systems in Containerization and Microservices through a Systematic Mapping. The study aimed to identify the primary methods, tools, and techniques used in the last five years to evaluate the performance of these systems. The results showed a lack of research in this area, with only four publications in the last year. However, the participation of research institutes and independent companies in these studies suggests a growing interest in the industry.

One significant gap identified was the limited use of analytical modeling to prepare precise performance estimation schemes. Additionally, the influence of communication protocols on performance analysis was not fully considered in some studies, which could affect the accuracy of the results. To address these gaps, future research should focus on developing analytical models that account for detailed aspects of Data Streaming systems operation, including communication protocols. This approach would lead to more precise performance analysis, which is crucial for the effective deployment of distributed systems.

Overall, the findings of this study support the hypothesis stated in the introduction that there are gaps in the performance analysis of Data Streaming systems in Containerization and Microservices. However, the research also highlights the need for further investigation to improve the accuracy and effectiveness of performance analysis techniques in this domain.

References

[1] J. Dizdarević, "Francisco carpio, admela jukan, and xavi masip-bruin. 2019. a survey of communication protocols for internet of things and related challenges of fog and cloud computing integration," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, pp. 1–29, 2019.

[2] R. Lu, G. Wu, B. Xie, and J. Hu, "Stream bench: Towards benchmarking modern distributed stream computing frameworks," in *2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*. IEEE, 2014, pp. 69–78.

[3] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and issues in data stream systems," in *Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2002, pp. 1–16.

[4] L. Querzoni and N. Rivetti, "Data streaming and its application to stream processing: tutorial," in *Proceedings of the 11th ACM International Conference on Distributed and Event-based Systems*, 2017, pp. 15–18.

[5] J. Dantas, E. Araujo, P. Maciel, R. Matos, and J. Teixeira, "Estimating capacity-oriented availability in cloud systems," *Int. Journal of Computational Science and Engineering*, vol. 22, no. 4, pp. 466–476, 2020.

[6] M. C. Bezerra, R. Melo, J. Dantas, and P. Maciel, "Availability evaluation of a vod streaming cloud service," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 765–770.

[7] B. A. Kitchenham, D. Budgen, and P. Brereton, *Evidence-based software engineering and systematic reviews*. CRC press, 2015, vol. 4.

[8] M. R. Henzinger, P. Raghavan, and S. Rajagopalan, "Computing on data streams." *External memory algorithms*, vol. 50, pp. 107–118, 1998.

[9] S. Keele *et al.*, "Guidelines for performing systematic literature reviews in software engineering," Technical report, ver. 2.3 ebse technical report. ebse, Tech. Rep., 2007.

[10] L. Wei, S. Hou, and Q. Liu, "Clinical care of hyperthyroidism using wearable medical devices in a medical iot scenario," *Journal of Healthcare Engineering*, vol. 2022, 2022.

[11] S. A. Noghabi, K. Paramasivam, Y. Pan, N. Ramesh, J. Bringhurst, I. Gupta, and R. H. Campbell, "Samza: stateful scalable stream processing at linkedin," *Proc. of the VLDB Endowment*, vol. 10, no. 12, pp. 1634–1645, 2017.

[12] G. Coviello, K. Rao, B. Debnath, O. Po, and S. Chakradhar, "Dataxe: A system for application self-optimization in serverless edge computing environments," in *Int. Conf. on Pervasive Computing and Communications Workshops (PerCom Workshops 2022)*. IEEE, 2022, pp. 699–705.

[13] C. Martín, P. Langendoerfer, P. S. Zarrin, M. Díaz, and B. Rubio, "Kafka-ml: Connecting the data stream with ml/ai frameworks," *Future Generation Computer Systems*, vol. 126, pp. 15–33, 2022.

[14] M. U. Ilyas, M. Ahmad, and S. Saleem, "Internet-of-things-infrastructure-as-a-service: The democratization of access to public internet-of-things infrastructure," *Int. Journal of Communication Systems*, vol. 33, no. 16, p. e4562, 2020.

[15] H. Wu, Z. Shang, and K. Wolter, "Trak: A testing tool for studying the reliability of data delivery in apache kafka," in *2019 IEEE International Symposium on Software Reliability Engineering Workshops (ISSREW)*. IEEE, 2019, pp. 394–397.

[16] A. Smid, R. Wang, and T. Cerny, "Case study on data communication in microservice architecture," in *Proceedings of the Conference on Research in Adaptive and Convergent Systems*, 2019, pp. 261–267.

[17] E. Badidi, "Towards a message broker based platform for real-time streaming of urban iot data," in *Computational and Statistical Methods in Intelligent Systems 2*. Springer, 2019, pp. 39–49.

[18] R. Scolati, I. Fronza, N. El Ioini, A. Samir, and C. Pahl, "A containerized big data streaming architecture for edge cloud computing on clustered single-board devices." in *Closer*, 2019, pp. 68–80.

[19] G. Hesse, C. Matthies, K. Glass, J. Huegle, and M. Uflacker, "Quantitative impact evaluation of an abstraction layer for data stream processing systems," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 1381–1392.

[20] N. Penchalaiah, A. S. Al-Humaimeedy, M. Maashi, J. C. Babu, O. I. Khalar, and T. H. Aldhyani, "Clustered single-board devices with docker container big stream processing architecture," *CMC-COMPUTERS MATERIALS & CONTINUA*, vol. 73, no. 3, pp. 5349–5365, 2022.

[21] I. Koren, "A standalone webassembly development environment for the internet of things," in *International Conference on Web Engineering*. Springer, 2021, pp. 353–360.

[22] F. Carpio, M. Delgado, and A. Jukan, "Engineering and experimentally benchmarking a container-based edge computing system," in *Int. Conf. on Communications (ICC 2020)*. IEEE, 2020, pp. 1–6.

[23] R. Scolati, I. Fronza, N. El Ioini, A. Samir, H. R. Barzegar, and C. Pahl, "A containerized edge cloud architecture for data stream processing," in *Cloud Computing and Services Science: 9th International Conference (CLOSER 2019)*. Springer, 2020, pp. 150–176.

[24] M. Gazzetti, A. Reale, K. Katrinis, and A. Corradi, "Scalable linux container provisioning in fog and edge computing platforms," in *Parallel Processing Workshops (Euro-Par 2017)*. Springer, 2018, pp. 304–315.

[25] K. S. Trivedi, D. S. Kim, A. Roy, and D. Medhi, "Dependability and security models," in *2009 7th International Workshop on Design of Reliable Communication Networks*. IEEE, 2009, pp. 11–20.