Optimizing Data Processing

Subjects: Computer Science, Information Systems

Contributor: Thanda Shwe , Masayoshi Aritsugi

Intelligent applications in several areas increasingly rely on big data solutions to improve their efficiency, but the processing and management of big data incur high costs. Although cloud-computing-based big data management and processing offer a promising solution to provide scalable and abundant resources, the current cloud-based big data management platforms do not properly address the high latency, privacy, and bandwidth consumption challenges that arise when sending large volumes of user data to the cloud. Computing in the edge and fog layers is quickly emerging as an extension of cloud computing used to reduce latency and bandwidth consumption, resulting in some of the processing tasks being performed in edge/fog-layer devices. Although these devices are resource-constrained, recent increases in resource capacity provide the potential for collaborative big data processing.

big data IoT serverless data processing

1. Introduction

In the last few years, a new class of applications that use unprecedented amounts of data generated from mobile devices and Internet of Things (IoT) sensors has been widely deployed in many areas, bringing better quality of life to humans through the automation of daily tasks and enabling time and energy savings, as well as monitoring, efficient communication, better decision making, etc. [1][2][3]. This type of application, which is present in various domains, such as healthcare monitoring, smart cities, industry, and transportation [4], not only involves performing intensive computation on large amounts of sensor data but also occasionally requires the output to be processed in real time to provide faster interactions and better user experiences. To provide the necessary computing resources, such as virtual machines and storage resources, including object-based storage and databases for these types of applications, cloud solutions are considered, as they are capable of storing and processing data in the cloud and sending the results back to the IoT devices [5][6]. Most existing systems use big data management platforms, which deal with large amounts of big data processing in the cloud computing platform or in self-hosted, dedicated clusters or data centers [7][8]. However, some IoT applications, such as smart cities and machinery automation, are real-time applications, and the adoption of cloud computing for such applications can result in high latency because sensor data need to be transferred to a cloud data center in a remote location. Thus, cloud solutions may no longer be appropriate for these types of applications due to the limitations caused by network bandwidth consumption and latency constraints [9].

Another noteworthy issue with cloud-based processing is that all users' data tend to be sent to the cloud. Specifically, there is potential for security and privacy issues and an increase in the communication payload and

costs. The storage of sensitive data in the cloud raises concerns about data vulnerability. In a centralized and shared environment, there is a risk of unauthorized access or data breaches. Regarding data privacy, the geographic locations of cloud servers may impact data privacy. Navigating international data protection laws and ensuring data residency compliance becomes a complex task in terms of protecting sensitive information. In addition, transferring large volumes of data to and from the cloud can incur significant costs. Organizations must carefully manage data transfer to reduce their expenses. These issues are particularly relevant in the context of data processing applications, such as healthcare and sensitive data handling, where the importance of security and low latency necessitates more focused consideration [10][11][12][13][14]. Concerns about sensitive information being exposed to remote cloud data centers highlight the limitations of cloud-based processing.

To overcome this challenge, fog and edge computing, which move the processing and storage tasks to locations that are closer to the data source, have been introduced as complements to cloud computing to provide storage and computing capabilities [9][15][16]. In these types of computing, any local device, however small, is considered capable of some processing tasks. Edge/fog computing helps users to perform data processing at lower latency and in the most appropriate way. At present, edge- and fog-layer devices are only used for some preprocessing tasks, such as filtering, feature processing, and compression. Computationally intensive tasks such as big data analytics and machine learning training are sent out to the cloud, as they cannot be efficiently run on low-performance local edge and fog devices [17][18]. Data processing in a layer architecture allows real-time and lightweight tasks processed on edge devices that need high storage capacities and processing capabilities to be moved to the cloud. However, important drawbacks related to the high bandwidth consumption involved in sending large quantities of data between local devices and cloud providers emerge. If the data can be stored and processed locally, closer to the IoT devices, the data transmitted through the network would be significantly reduced.

Researchers can observe a recent breakthrough in edge- and fog-layer computing devices as the processing abilities of these devices have become increasingly efficient and capable [16]. In addition, most of the data or applications pass through the edge/fog-layer devices as a gateway before connecting to the cloud. Edge/fog computing is potentially a good solution for data processing problems, thanks to its recently developed resource-scaling capabilities. This provides the possibility of deploying edge- and fog-layer big data management platforms using local edge/fog devices as computing and storage nodes. Such a big data management platform can not only help to reduce latency, bandwidth consumption, and cloud budgets but can also be applied in some environments without an Internet connection.

IoT applications that span various domains, ranging from smart cities to industrial environments and healthcare systems, continuously produce massive volumes of data, including sensor readings, images, videos, and other types of information. Data processing platforms handle large volumes of diverse data, making them well-suited for the processing of the varied data generated by IoT sources. IoT deployments often involve distributed environments with numerous devices spread across different locations. The distributed computing models of big data processing allow them to efficiently process data across distributed clusters, accommodating the distributed nature of IoT deployments. As many IoT applications demand real-time or near-real-time processing to enable swift

responses and decision making, centralized data management approaches encounter challenges in the efficient handling of information. Thus, by strategically integrating computational capabilities closer to the data source, within the edge and fog layers, it is possible to facilitate local data processing, reducing the necessity of transmitting all data to centralized cloud infrastructure. This not only addresses latency concerns but also optimizes bandwidth usage, making it particularly beneficial in scenarios where network resources are limited. Furthermore, the decentralized nature of edge/fog computing contributes to enhanced security and privacy by allowing sensitive data to be processed closer to their origin. In summary, the cooperative interconnection between IoT applications, data management, and edge/fog computing represents a strategic approach to meet the evolving demands of efficient, real-time, and secure data processing within the dynamic landscape of the Internet of Things.

In addition, the requirements of big data processing, such as the training of machine learning models [19][20][21], stream processing [22], and event processing [23][24], raise scalability issues and require significant computing and storage resources [25][26]. Fortunately, big data management and processing platforms such as Apache Spark [27], Apache Flink [28], and Apache OpenWhisk [29] address these concerns by facilitating distribution and collaboration. In such systems, big data processing tasks are performed using the collaborative power of nodes in clusters. With recent increases in the resource capacity of edge/fog devices and the capability of big data processing platforms, it has become increasingly important to explore the deployment of big data processing platforms on resource-constrained edge/fog devices.

2. Optimizing Data Processing

2.1. Capability of Resource-Constrained Devices

Due to the increasing capabilities of low-cost, single-board computers and their benefits in terms of cost, low power consumption, small size, and reasonable performance, several works [30][31][32] have studied the capabilities and performance of low-cost edge-layer devices for deep learning and machine learning interference. The works reported in [30][31] focused on the inference of deep learning models in various low-cost devices, such as the Raspberry Pi 4, Google Coral Dev Board, and Nvidia Jetson Nano, and evaluated their performance in terms of the inference time and power consumption. Their main focus was to implement the inference parts of deep learning models in edge devices in order to achieve real-time processing. The authors of [32] conducted a comprehensive survey of design methodologies for AI edge development, emphasizing the importance of single-layer specialization and cross-layer codesign, which includes hardware and software components for edge training, inference, caching, and offloading. Several insights were highlighted with respect to the quality of AI solutions in the edge computing layer. While the deployment of low-cost edge devices is still a relatively new paradigm for advanced applications, it has found widespread interest, particularly in function processing. Some approaches [33][34][35][36][37] have been developed for edge-layer serverless platforms. Specifically, all of these works attempted to develop edge-layer serverless platforms that can support edge AI applications. Moreover, increasing attention has been paid to the capabilities and viability of single-board computers; several works [38][39][40] have explored the characteristics and possibilities of deploying different big data applications in resource-constrained environments.

Some previous studies have evaluated data processing platforms under different application scenarios and proposed design changes to off-the-shelf software platforms to cater to the requirements of applications in edge/fog computing layers. In [41], the authors proposed a serverless edge platform based on the OpenWhisk serverless platform to address the challenges of real-time and data-intensive applications in fog/edge computing. The proposed platform comprised additional components for latency-sensitive computation offloading, stateful partitions, and real-time edge coordination. The platform was evaluated in terms of its resource utilization footprint, latency overhead, throughput, and scalability under different application scenarios. The results show that the serverless architecture reduced the burden of infrastructure management, allowed greater functionality to be deployed on fog nodes with limited resources, and fulfilled the requirements of different application scenarios and the heterogeneous deployment of fog nodes. To optimize stream processing in the edge/fog environment, the authors of [42] proposed Amnis, a stream processing framework that considers computational and network resources at the edge. It extended the Storm framework in terms of stream processing operator allocation and placement for stream queries. Compared to the default operator scheduler in Apache Storm, it performed better in terms of end-to-end latency and overall throughput. These works targeted only individual data processing paradigms, such as batch processing, stream processing, or function processing.

2.3. Comparative Studies of Big Data Processing Platforms

Some previous works have included benchmark studies of the performance of stream processing systems, such as [43], which evaluated the performance of three stream processing platforms, namely Apache Storm, Apache Spark, and Apache Flink, in terms of throughput and latency. In [44], these aspects were also evaluated for Apache Spark and Apache Flink. With respect to batch processing platforms, the work reported in [44][45] included a comprehensive study of two widely used big data analytics tools, namely Apache Spark [27] and Hadoop MapReduce [46], on a common data mining task, i.e., classification. With respect to function processing platforms, the work reported in [47][48] provided a benchmarking framework for the characterization of serverless platforms in both commercial cloud and open-source platforms. The work that is most related to ours is [49], which evaluated computing resources across the computing continuum using three applications: video encoding, machine learning, and in-memory analytics. It also provided recommendations based on the evaluation results regarding where to perform the tasks across the computing continuum. The authors utilized a real test bed named the Carinthian Computing Continuum (C 3) to extend cloud data centers with low-cost devices located close to the edge of the network. They recommended offloading the applications to edge and fog resources to reduce the network traffic and CO22 emissions, with an acceptable performance penalty, while the cloud was used for lower execution times. Another work [50] mainly evaluated big data processing possibilities on a Raspberry Pi-based Apache Spark and Hadoop Cluster and examined the impact on storage performance of employing three different external storage solutions. Then, the cluster performance was compared with that of a single desktop PC using microbenchmarks such as Wordcount, TeraGen/TeraSort, TestDFSIO, and Pi computation.

However, neither of these works included a comprehensive investigation of the performance of existing big data processing platforms for the three computing paradigms, namely batch processing, stream processing, and function processing, based on various applications (e.g., image classification, object detection, image resizing).

Therefore, exploring the operating performance of batch, stream, and function processing in the current edge, fog, and cloud layers is of great significance for research and industrial applications in the field of big data processing.

References

1. Rani, S.; Chauhan, M.; Kataria, A.; Khang, A. IoT Equipped Intelligent Distributed Framework for Smart Healthcare Systems. In Towards the Integration of IoT, Cloud and Big Data: Services, Applications and Standards; Rishiwal, V., Kumar, P., Tomar, A., Malarvizhi Kumar, P., Eds.; Springer Nature Singapore: Singapore, 2023; pp. 97–114.
2. Sarker, I.H.; Khan, A.I.; Abushark, Y.B.; Alsolami, F. Internet of Things (IoT) Security Intelligence: A Comprehensive Overview, Machine Learning Solutions and Research Directions. Mob. Netw. Appl. 2023, 28, 296–312.
3. Mukati, N.; Namdev, N.; Dilip, R.; Hemalatha, N.; Dhiman, V.; Sahu, B. Healthcare Assistance to COVID-19 Patient using Internet of Things (IoT) Enabled Technologies. Mater. Today Proc. 2023, 80, 3777–3781.
4. The Top 10 IoT Segments in 2018—Based on 1,600 Real IoT Projects. Available online: https://iot-analytics.com/top-10-iot-segments-2018-real-iot-projects/ (accessed on 20 July 2023).
5. Alhaidari, F.; Rahman, A.; Zagrouba, R. Cloud of Things: Architecture, applications and challenges. J. Ambient. Intell. Humaniz. Comput. 2023, 14, 5957–5975.
6. Turukmane, A.V.; Pradeepa, M.; Reddy, K.S.S.; Suganthi, R.; Riyazuddin, Y.; Tallapragada, V.S. Smart farming using cloud-based Iot data analytics. Meas. Sens. 2023, 27, 100806.
7. Alam, T. Cloud-Based IoT Applications and Their Roles in Smart Cities. Smart Cities 2021, 4, 1196–1219.
8. Rajabion, L.; Shaltooki, A.A.; Taghikhah, M.; Ghasemi, A.; Badfar, A. Healthcare big data processing mechanisms: The role of cloud computing. Int. J. Inf. Manag. 2019, 49, 271–289.
9. Bonomi, F.; Milito, R.; Zhu, J.; Addepalli, S. Fog Computing and Its Role in the Internet of Things. In Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, MCC ’12, New York, NY, USA, 17 August 2012; pp. 13–16.
10. Johri, P.; Balu, V.; Jayaprakash, B.; Jain, A.; Thacker, C.; Kumari, A. Quality of service-based machine learning in fog computing networks for e-healthcare services with data storage system. Soft Comput. 2023.
11. Azizi, S.; Farzin, P.; Shojafar, M.; Rana, O. A scalable and flexible platform for service placement in multi-fog and multi-cloud environments. J. Supercomput. 2023.

12. Karatas, M.; Eriskin, L.; Deveci, M.; Pamucar, D.; Garg, H. Big Data for Healthcare Industry 4.0: Applications, challenges and future perspectives. Expert Syst. Appl. 2022, 200, 116912.
13. Agapito, G.; Cannataro, M. An Overview on the Challenges and Limitations Using Cloud Computing in Healthcare Corporations. Big Data Cogn. Comput. 2023, 7, 68.
14. Quy, V.K.; Hau, N.V.; Anh, D.V.; Ngoc, L.A. Smart healthcare IoT applications based on fog computing: Architecture, applications and challenges. Complex Intell. Syst. 2022, 8, 3805–3815.
15. Yi, S.; Hao, Z.; Qin, Z.; Li, Q. Fog Computing: Platform and Applications. In Proceedings of the 2015 Third IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb), Washington, DC, USA, 12–13 November 2015; pp. 73–78.
16. Muniswamaiah, M.; Agerwala, T.; Tappert, C.C. Fog Computing and the Internet of Things (IoT): A Review. In Proceedings of the 2021 8th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2021 7th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), Washington, DC, USA, 26–28 June 2021; pp. 10–12.
17. Saini, K.; Kalra, S.; Sood, S.K. An Integrated Framework for Smart Earthquake Prediction: IoT, Fog, and Cloud Computing. J. Grid Comput. 2022, 20, 17.
18. Verma, P.; Tiwari, R.; Hong, W.C.; Upadhyay, S.; Yeh, Y.H. FETCH: A Deep Learning-Based Fog Computing and IoT Integrated Environment for Healthcare Monitoring and Diagnosis. IEEE Access 2022, 10, 12548–12563.
19. Alazzam, H.; AbuAlghanam, O.; Sharieh, A. Best path in mountain environment based on parallel A* algorithm and Apache Spark. J. Supercomput. 2022, 78, 5075–5094.
20. Bagui, S.; Walauskis, M.; DeRush, R.; Praviset, H.; Boucugnani, S. Spark Configurations to Optimize Decision Tree Classification on UNSW-NB15. Big Data Cogn. Comput. 2022, 6, 38.
21. Chebbi, I.; Mellouli, N.; Farah, I.R.; Lamolle, M. Big Remote Sensing Image Classification Based on Deep Learning Extraction Features and Distributed Spark Frameworks. Big Data Cogn. Comput. 2021, 5, 21.
22. Ed-daoudy, A.; Maalmi, K.; El Ouaazizi, A. A scalable and real-time system for disease prediction using big data processing. Multimed. Tools Appl. 2023, 82, 30405–30434.
23. Kumari, A.; Behera, R.K.; Sahoo, B.; Misra, S. Role of Serverless Computing in Healthcare Systems: Case Studies. In Proceedings of the Computational Science and Its Applications—ICCSA 2022 Workshops; Gervasi, O., Murgante, B., Misra, S., Rocha, A.M.A.C., Garau, C., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 123–134.
24. Bac, T.P.; Tran, M.N.; Kim, Y. Serverless Computing Approach for Deploying Machine Learning Applications in Edge Layer. In Proceedings of the 2022 International Conference on Information Networking (ICOIN), Jeju-si, Republic of Korea, 12–15 January 2022; pp. 396–401.

25. Kong, L.; Tan, J.; Huang, J.; Chen, G.; Wang, S.; Jin, X.; Zeng, P.; Khan, M.; Das, S.K. Edge-Computing-Driven Internet of Things: A Survey. ACM Comput. Surv. 2022, 55, 1–44.
26. Singh, R.; Gill, S.S. Edge AI: A survey. Internet Things Cyber-Phys. Syst. 2023, 3, 71–92.
27. Apache Spark. Available online: https://spark.apache.org/ (accessed on 8 August 2023).
28. Apache Flink. Available online: https://flink.apache.org/ (accessed on 8 August 2023).
29. Apache OpenWhisk. Available online: https://openwhisk.apache.org/ (accessed on 8 August 2023).
30. Baller, S.P.; Jindal, A.; Chadha, M.; Gerndt, M. DeepEdgeBench: Benchmarking Deep Neural Networks on Edge Devices. In Proceedings of the 2021 IEEE International Conference on Cloud Engineering (IC2E), San Francisco, CA, USA, 4–8 October 2021; pp. 20–30.
31. Feng, H.; Mu, G.; Zhong, S.; Zhang, P.; Yuan, T. Benchmark Analysis of YOLO Performance on Edge Intelligence Devices. Cryptography 2022, 6, 16.
32. Hao, C.; Dotzel, J.; Xiong, J.; Benini, L.; Zhang, Z.; Chen, D. Enabling Design Methodologies and Future Trends for Edge AI: Specialization and Codesign. IEEE Des. Test 2021, 38, 7–26.
33. Rausch, T.; Hummer, W.; Muthusamy, V.; Rashed, A.; Dustdar, S. Towards a Serverless Platform for Edge AI. In Proceedings of the 2nd USENIX Workshop on Hot Topics in Edge Computing (HotEdge 19), Renton, WA, USA, 9 July 2019.
34. Pfandzelter, T.; Bermbach, D. tinyFaaS: A Lightweight FaaS Platform for Edge Environments. In Proceedings of the 2020 IEEE International Conference on Fog Computing (ICFC), Sydney, NSW, Australia, 21–24 April 2020; pp. 17–24.
35. Smith, C.P.; Jindal, A.; Chadha, M.; Gerndt, M.; Benedict, S. FaDO: FaaS Functions and Data Orchestrator for Multiple Serverless Edge-Cloud Clusters. In Proceedings of the 2022 IEEE 6th International Conference on Fog and Edge Computing (ICFEC), Messina, Italy, 16–19 May 2022; pp. 17–25.
36. Großmann, M.; Ioannidis, C.; Le, D.T. Applicability of Serverless Computing in Fog Computing Environments for IoT Scenarios. In Proceedings of the 12th IEEE/ACM International Conference on Utility and Cloud Computing Companion, UCC ’19 Companion, Auckland, New Zealand, 2–5 December 2019; pp. 29–34.
37. Tzenetopoulos, A.; Apostolakis, E.; Tzomaka, A.; Papakostopoulos, C.; Stavrakakis, K.; Katsaragakis, M.; Oroutzoglou, I.; Masouros, D.; Xydis, S.; Soudris, D. FaaS and Curious: Performance Implications of Serverless Functions on Edge Computing Platforms. In Proceedings of the High Performance Computing: ISC High Performance Digital 2021 International Workshops, Frankfurt am Main, Germany, June 24–July 2, 2021, Revised Selected Papers 36;

Jagode, H., Anzt, H., Ltaief, H., Luszczek, P., Eds.; Springer International Publishing: Cham, Switzerland, 2021; pp. 428–438.

38. Ab Wahab, M.N.; Nazir, A.; Zhen Ren, A.T.; Mohd Noor, M.H.; Akbar, M.F.; Mohamed, A.S.A. Efficientnet-Lite and Hybrid CNN-KNN Implementation for Facial Expression Recognition on Raspberry Pi. IEEE Access 2021, 9, 134065–134080.
39. James, N.; Ong, L.Y.; Leow, M.C. Exploring Distributed Deep Learning Inference Using Raspberry Pi Spark Cluster. Future Internet 2022, 14, 220.
40. Curtin, B.H.; Matthews, S.J. Deep Learning for Inexpensive Image Classification of Wildlife on the Raspberry Pi. In Proceedings of the 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, 10–12 October 2019; pp. 82–87.
41. Baresi, L.; Filgueira Mendonça, D. Towards a Serverless Platform for Edge Computing. In Proceedings of the 2019 IEEE International Conference on Fog Computing (ICFC), Prague, Czech Republic, 24–26 June 2019; pp. 1–10.
42. Xu, J.; Palanisamy, B.; Wang, Q.; Ludwig, H.; Gopisetty, S. Amnis: Optimized stream processing for edge computing. J. Parallel Distrib. Comput. 2022, 160, 49–64.
43. Karimov, J.; Rabl, T.; Katsifodimos, A.; Samarev, R.; Heiskanen, H.; Markl, V. Benchmarking Distributed Stream Data Processing Systems. In Proceedings of the 2018 IEEE 34th International Conference on Data Engineering (ICDE), Paris, France, 16–19 April 2018; pp. 1507–1518.
44. Veiga, J.; Expósito, R.R.; Pardo, X.C.; Taboada, G.L.; Tourifio, J. Performance evaluation of big data frameworks for large-scale data analytics. In Proceedings of the 2016 IEEE International Conference on Big Data (Big Data), Washington, DC, USA, 5–8 December 2016; pp. 424–431.
45. Tekdogan, T.; Cakmak, A. Benchmarking Apache Spark and Hadoop MapReduce on Big Data Classification. In Proceedings of the 2021 5th International Conference on Cloud and Big Data Computing, ICCBDC ’21, New York, NY, USA, 17–18 March 2021; pp. 15–20.
46. Apache Hadoop. Available online: https://hadoop.apache.org/ (accessed on 8 August 2023).
47. Grambow, M.; Pfandzelter, T.; Burchard, L.; Schubert, C.; Zhao, M.; Bermbach, D. BeFaaS: An Application-Centric Benchmarking Framework for FaaS Platforms. In Proceedings of the 2021 IEEE International Conference on Cloud Engineering (IC2E), San Francisco, CA, USA, 4–8 October 2021; pp. 1–8.
48. Yu, T.; Liu, Q.; Du, D.; Xia, Y.; Zang, B.; Lu, Z.; Yang, P.; Qin, C.; Chen, H. Characterizing Serverless Platforms with Serverlessbench. In Proceedings of the 11th ACM Symposium on Cloud Computing, SoCC ’20, New York, NY, USA, 19–21 October 2020; pp. 30–44.

49. Kimovski, D.; Mathá, R.; Hammer, J.; Mehran, N.; Hellwagner, H.; Prodan, R. Cloud, Fog, or Edge: Where to Compute? IEEE Internet Comput. 2021, 25, 30–36.
50. Lee, E.; Oh, H.; Park, D. Big Data Processing on Single Board Computer Clusters: Exploring Challenges and Possibilities. IEEE Access 2021, 9, 142551–142565.