Scalable and Dynamic Big Data Processing and Service Provision in Edge Cloud Environments

In-Young Ko[1], Abhishek Srivastava[2] and Michael Mrissa[3]

[1]*Korea Advanced Institute of Science and Technology, South Korea*
[2]*Indian Institute of Technology Indore, India*
[3]*InnoRenew CoE, University of Primorska, Slovenia*
E-mail: iko@kaist.ac.kr; asrivastava@iiti.ac.in; michael.mrissa@innorenew.eu

Owing to the exponential growth of connected devices and the large amounts of data produced by such devices, clouds are becoming a bottleneck and cause latency while collecting and processing data and providing associated services [1]. The concept of *edge computing* has been suggested to solve this scalability problem by moving data centers and computing resources close to the data sources [2]. Locally deployed data centers and computing resources form an *edge cloud* or a *fog* that can collect and process big data in a distributed and scalable manner [3]. Recently, low-latency and reliable communication technologies such as 5G have enabled more effective realization of edge cloud environments [4].

Edge clouds are especially useful for efficient, reliable, and secure data collection and processing for smart cities and smart factories [5]. Figure 1 shows an overview of big data-driven service provision in a smart city environment. In this environment, edge devices such as cars and sensors on roads collect data and accumulate them in nearby edge clouds. Then, the services running on an edge cloud process them locally and provide localized service capabilities to nearby users. Such services can be dynamically deployed to edge clouds according to user needs and various environmental contexts.

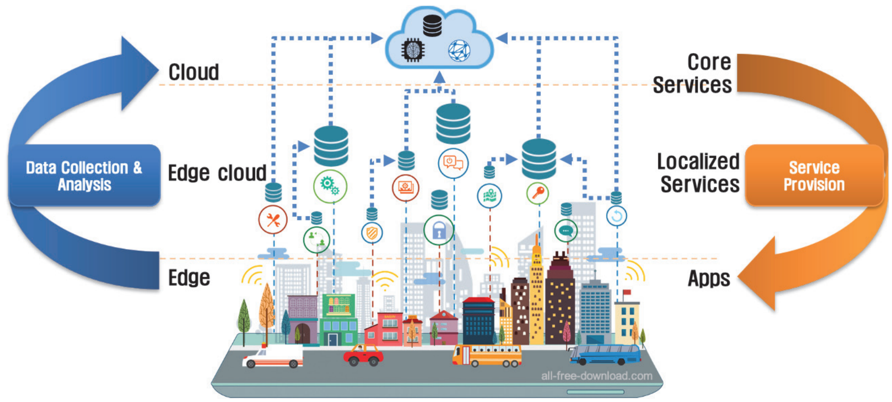

Figure 1 Overview of big data-driven service provision in an edge cloud environment.

Therefore, in edge cloud environments, it is essential to provide services to efficiently collect and process various types of big data in real time. In addition, it is necessary to build a framework for developing value-adding big data-driven applications in a reliable and usable manner by utilizing the services. The first international workshop on big data-driven edge cloud services (BECS 2021)[1] was held to provide a venue where scholars and practitioners can share their experiences and present ongoing work on these issues. The workshop was held in conjunction with the 21st International Conference on Web Engineering (ICWE 2021),[2] which was held online on May 18–21, 2021.

The first edition of the BECS workshop was specially focused on the following topics: Web services in edge clouds, Web of Things in edge clouds, AI in edge computing (Edge AI), dependable and highly usable big data platforms, distributed data collection, analysis and prediction, stream data processing in edge clouds, big knowledge graphs for distributed edge cloud environments, modeling and mashup of edge cloud services, micro-service architecture for edge cloud environments, and edge-cloud interaction and collaboration.

This special issue of the *Journal of Web Engineering* is oriented toward discussing the aforementioned topics dealing with selected papers from the BECS 2021 workshop.

---

[1]https://becs.kaist.ac.kr/iwbecs2021/

[2]https://icwe2021.webengineering.org/

Vargas-Solar et al. introduced an innovative composable just-in-time architecture for configuring data centers for data science pipelines (JITA-4DS) and associated resource management techniques for more dynamic and flexible big data processing in edge cloud environments. Most existing data science (DS) pipelines provide one-fits-all solutions, imposing the complete externalization of data pipeline tasks. However, in this approach, the authors distribute some of the DS tasks across edges to provide just-in-time resources while ensuring ad hoc and elastic execution environments. JITA-4DS is a cross-layer management system that is aware of both the application characteristics and the underlying infrastructures to break the barriers between applications, middleware/operating systems, and hardware layers. Vertical integration of these layers is required for building a customizable virtual data center to meet the requirements of dynamically changing data science pipelines, such as performance, availability, and energy consumption. The authors presented an experimental simulation devoted to running data science workloads and determining the best strategies for scheduling the allocation of resources implemented by JITA-4DS.

Service recommendation based on quality of service (QoS) has been an important issue in Web service computing. Recently, to improve the accuracy of service recommendation, researchers have applied machine learning techniques. However, unlike traditional Web services, there are various factors that affect the quality of services that are provided via distributed edge clouds, and therefore, they make accurate QoS prediction and service recommendation more challenging. Choi et al. proposed an approach called GAIN-QoS to simulate a real-world edge computing environment and to improve the QoS prediction performance in edge cloud environments. It clusters services based on their location, calculates the distance between services and users in each cluster, and brings the QoS values of users within a certain distance. The authors applied a generative adversarial imputation (GAIN) model and performed QoS prediction on the basis of this reconstructed user service invocation matrix. By conducting an experiment, the authors showed that the proposed approach can significantly improve the accuracy of QoS prediction for edge cloud environments, which suffer from the cold-start problem.

Rentero-Trejo et al. dealt with edge cloud environments where there are many Internet-of-Things (IoT) devices connected to provide services to nearby users. The authors proposed a novel solution based on federated learning, where IoT devices play a more active role as edge nodes for learning the preferences of their owners and adapting the behavior of already known and new smart environments to these preferences. To obtain context-aware and personalized predictions, the authors proposed two models: a global model with the knowledge generated in the federation and that provides predictions to new users and/or new environments, and a personalized model that adjusts to the needs of a particular user for already known environments. The authors showed that the proposed approach allows fast personalization and deployment, offering predictions in every environment the user visits and managing multiple environments.

On the other hand, Hrovatin et al. considered a sensor network deployed to detect events using a convolutional neural network (CNN) that uses the input from multiple sensor nodes. Because the CNN inference process requires data from multiple sensor nodes, typical solutions convey all the data of the nodes to a central processing point, which is usually in the cloud. Therefore, transferring data to the cloud could introduce security concerns and substantial latency between event occurrence and detection. To tackle these problems, the authors proposed an approach for distributing the CNN processing load across the actual data sources of the sensor nodes of the network. More specifically, they proposed a technique applicable to grid-shaped sensor networks, in which neighboring sensor nodes interchange sensor readings to conjointly compute the convolution layers of a CNN on sensor nodes. Additionally, they discussed the application of the proposed technique to their nonintrusive fall detection sensor network, dubbed the smart floor. The smart floor is a grid-shaped sensor network in which each sensor node senses the local force applied to the floor. A CNN is used to recognize whether the activity occurring over the smart floor is a person that fell on the floor or just activities of daily living, such as walking and moving objects.

Finally, Tošić and Vičič showed the potential use of geosharding in decentralized routing networks to improve fault tolerance and latency in edge cloud environments. Such networks can be used as a communication layer for edge devices that compute large amounts of data. Specifically, the authors focused on the Low Latency Anonymous Routing Protocol (LLARP), a protocol built on top of the Oxen blockchain that aims to achieve Internet privacy. They analyzed the existing network of service nodes, observed cloud provider centralization, and proposed a high-level protocol that provides incentives for a better geographical distribution, mitigating potential cloud provider/country-wide service dropouts. Additionally, the protocol level information about geographical location can be used to improve the client's path selection and decrease network latency. They demonstrated the feasibility of their approach by comparing it with the random path selection in a simulated environment.

They observed marginal drops in the average latency when selecting paths geographically closer to each other.

Acknowledgment

The BECS 2021 workshop was supported by the Ministry of Science and ICT (MSIT), Korea, under the Information Technology Research Center (ITRC) support program (IITP-2021-2020-0-01795) supervised by the Institute of Information & Communications Technology Planning Evaluation (IITP). It was also supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2019R1A2C1087430).

References

[1] Shaukat Ali, Ferruccio Damiani, Schahram Dustdar, Marialuisa Sanseverino, Mirko Viroli, and Danny Weyns. Big data from the cloud to the edge: the aggregate computing solution. In *Proceedings of the 13th European Conference on Software Architecture – Volume 2 (ECSA '19)*, Association for Computing Machinery, New York, NY, USA, 177–180, 2019.

[2] Ejaz Ahmed, Arif Ahmed, Ibrar Yaqoob, Junaid Shuja, Abdullah Gani, Muhammad Imran, and Muhammad Shoaib. Bringing computation closer toward the user network: Is edge computing the solution? *IEEE Communications Magazine*, 55(11):138–144, 2017.

[3] Ashkan Yousefpour, Caleb Fung, Tam Nguyen, Krishna Kadiyala, Fatemeh Jalali, Amirreza Niakanlahiji, Jian Kong, Jason P. Jue. All one needs to know about fog computing and related edge computing paradigms: A complete survey. *Journal of Systems Architecture*, Volume 98, Pages 289–330, 2019.

[4] Anselme Ndikumana, Nguyen H. Tran, Tai Manh Ho, Zhu Han, Walid Saad, Dusit Niyato, and Choong Seon Hong. Joint Communication, Computation, Caching, and Control in Big Data Multi-Access Edge Computing *IEEE Transactions on Mobile Computing*, 19(6):1359–1374, 2020.

[5] Nouha Kherraf, Sanaa Sharafeddine, Chadi M. Assi, and Ali Ghrayeb. Latency and Reliability-Aware Workload Assignment in IoT Networks With Mobile Edge Clouds. *IEEE Transactions on Network and Service Management*, 16(4):1435–1449, 2019.