Data Lake Conceptualized Web Platform for Food Research Data Collection

Gi-taek An[1,2], Seyoung Oh[1], Eunhye Kim[1]
and Jung-min Park[1,*]

[1] *Korea Food Research Institute, Wanju-gun 55365, Republic of Korea*
[2] *Division of Computer Science and Artificial Intelligence, Jeonbuk National University, Jeonju 54896, Republic of Korea*
E-mail: gt@kfri.re.kr; ohsy@kfri.re.kr; eunhye@kfri.re.kr; parkjm@kfri.re.kr
[*] *Corresponding Author*

Received 05 December 2023; Accepted 05 April 2024

Abstract

Food research is uniquely intertwined with everyday life and necessitates the utilization of big data. Within this domain, the research data consist of various forms and formats, encompassing biological experiment results, chemical analysis data, nutritional information, microbiological data, sensor data, images, and videos. This diversity stems from the integration of data from various subdomains within the larger field. With recent advancements in deep learning technology, the importance of data has grown significantly, resulting in increased reliance on data-driven research. Although specialized platforms for sharing and utilizing data have been established at the national level, particularly in the bioscience field, food research lacks a dedicated infrastructure and specialized data-sharing platforms. In this study, we develop a platform that leverages Hadoop-based distributed file systems to create a data lake. This platform enables data storage and sharing through a web-based interface. The distributed file system supports scalability by adding data nodes, making it an effective solution for capacity expansion. In addition, the web-based

Journal of Web Engineering, Vol. 23_3, 377–392.
doi: 10.13052/jwe1540-9589.2333

platform ensures high accessibility, allowing users access from anywhere, at any time, using any device. Finally, we introduce the establishment of a 1.8 PB Hadoop-based physical storage system and present an approach for building a highly accessible web platform with substantial utility.

Keywords: Food research data, big data, data platform, data collection, web-based platform.

1 Introduction

Food research is a crucial area that is closely related to daily life and health [1]. Research data in this field consist of various forms, such as biological experimental results, chemical analysis data, nutritional information, microbiological data, sensor data, images, and videos [2]. The interdisciplinary nature of food research often results in the integration of data from various domains. At present, the field of food research lacks dedicated infrastructure for storing, sharing, and utilizing research data, considering the specific requirements of the field for data sharing and utilization. Researchers struggle to share their experimental results and data, hindering their collaboration with their peers. Consequently, redundant research efforts increase, leading to reduced research efficiency. Furthermore, the COVID-19 pandemic rapidly transformed work patterns in both industrial and research settings. Remote working has become prevalent, and researchers now need to perform data analysis in locations other than laboratories. Given the growing importance of remote working owing to COVID-19, efficient methods for collecting, managing, and sharing research data have become even more critical.

In the field of bioscience, solutions to these issues have been found in data-sharing platforms. In particular, the United States, the European Union, Japan, and China have implemented large-scale projects and policies at the national level to promote bio-data sharing and collaboration [3–6]. These efforts are also expanding in the Republic of Korea [7]. However, an effective infrastructure for research data sharing is still lacking in the field of food research, thereby impeding innovation and progress in this area.

In this study, we address these challenges by developing a suitable data platform tailored to the forms and formats of the data that are encountered in the field of food research. Our data platform allows researchers to create a data management plan (DMP) and upload data with ease using the metadata that are generated in the plans. A data management plan is a document that outlines the lifecycle management of data collected or produced during the research process. The metadata can then be used for information retrieval and data sharing. Through this data platform, we aim to collect and share the data generated in various food research areas effectively, ultimately enhancing research efficiency and quality. In Section 2, we analyze case studies of platform construction and operation for research data. Section 3 explains the method for building a data platform that is suitable for the food industry. In Section 4, we discuss the limitations of this study and future research directions, and Section 5 describes the results of building a web-based food research data platform.

2 Related Work

One of the most popular approaches for creating a platform for collecting and utilizing data is to employ the concept of a data lake.

The first step in constructing a data platform is creating a data lake. The method for building a data lake typically involves utilizing the Hadoop ecosystem, which is open-source software that provides distributed processing of massive datasets across computer clusters while providing high availability [8]. Additionally, computer clusters can be scaled from a single server to thousands of computers. This distributed storage technology has applications in fields where large volumes of data are generated and real-time storage is required, such as in the use of sensor data and the storage and analysis of large-scale data in biosciences. Studies on the storage and processing of large datasets have been conducted in various fields. Examples include research on storing large-scale meteorological data [9], storing remotely sensed data that are produced in large quantities [10], using Hadoop for storing resource description framework models for linked data and knowledge graphs [11, 12], and research on using Hadoop for processing medical data to predict chronic kidney diseases in the context of large-scale bioscience data [13]. Other studies have explored the advantages of Hadoop in storing and searching proteomic datasets [14], as well as storing large-scale genomic data in FASTA/Q files [15]. In addition, research has been conducted to explore the use of Hadoop in connecting painting resources with a storage-and-retrieval system [16] and building data repositories by investigating storage formats and technologies for large-scale data [17].

The advantage of Hadoop in terms of accommodating various data formats without constraints makes it suitable for a wide range of data types, from massive sensor data to image and large-scale bioscience data. Given the diverse applications of food research in various fields, using the Hadoop file system to store food research data is an efficient option.

The second step is to configure a user interface for storing and utilizing the data. Even if a data lake is well constructed, if it is difficult for researchers to register the data and security measures are lacking, it cannot be used effectively. A unified data analysis application was created using modern computing infrastructure and web-based service platforms driven by software [18]. This application was designed in a web-based environment to facilitate community-driven data access and interaction among analysis platforms within a cloud environment. The implementation of web-based user interfaces has proven to be effective in enhancing interoperability and accessibility. Additionally, some studies have aimed to provide convenience in water quality research by effectively integrating the input, analysis, and output within web-based platforms [19], as well as offering online spaces for the real-time investigation and analysis of monitoring tools [20]. Certain studies have also opted for a web-based approach for visualizing and analyzing omics data [21–24]. In addition, due to COVID-19, there is active research and utilization of big data in the field of bioresearch [25–29]. With the recent availability of network-enabled devices that are equipped with web browsers as a standard feature, configuring the web as a user interface allows for consistent platform functionality across different devices. In the field of food research, unlike related studies, there is currently no dedicated platform for sharing and utilizing data. Thus, researchers have resorted to using data platforms from other research domains. In this study, we construct an infrastructure based on Hadoop to serve as a data lake capable of handling various data formats in food research. In addition, we develop a web-based platform that can manage both metadata and actual research data.

3 Implementation of a Web-based Data Platform

Two essential components are required for the construction of a data platform: the physical storage and the user interface, which allows users to store, access, and manage data. Owing to the diverse forms and formats of research data in the field of food research, the data repository should be built considering these characteristics. Additionally, the research data may be the results obtained from experiments or those used in ongoing experiments and should be free from damage due to system failures.

First, we describe the storage construction method using the Hadoop Distributed File System for the storage of data used in food research. This system allows for the flexible storage of a wide range of data, from small-capacity structured data in tabular format, which are commonly used in food research, to large-scale unstructured data such as images and photos. Furthermore, the files are distributed and replicated into specific block sizes, thereby avoiding data loss owing to system failures.

Second, we describe the web-based platform as an interface connected to the storage, which enables users to store, utilize, and access research data. The advancement of remote technologies and changes in corporate work patterns due to COVID-19 have also affected the research environment [30]. Analyzing and conducting research based on data are no longer limited to the laboratories. In a research environment where one can access research data anytime and anywhere, provided that access to data and analytical tools are available, that location can become a laboratory. Such an environment can be established using web technology.

A data platform based on web technology enables access and usage of data without being constrained by time, location, or devices. Moreover, it offers high interoperability with other data analysis platforms, making it convenient for various additional applications. This section outlines the method of constructing a data repository using distributed storage technology and building the corresponding web-based research data platform.

3.1 Infrastructure Establishment Using the Hadoop Distributed Storage System

Research data constitute the output of research and provide vital information for research. These research data represent significant achievements and resources for researchers. Therefore, when managing research data, researchers must ensure that the data are not compromised. Measures such as managing replicas based on structural changes in the data or regular backups at critical points are generally implemented. These management practices are also effective when building the infrastructure for storing research data. The storage that is used for retaining research data must ensure data integrity even in the face of system failures. In this study, we establish a data storage infrastructure using Hadoop-based distributed file storage, making it easily deployable and usable for a range of entities, from small-scale food companies and university research laboratories to large corporations and research institutions.

The construction of the distributed storage infrastructure using Hadoop is carried out in the following sequence: physical configuration, network

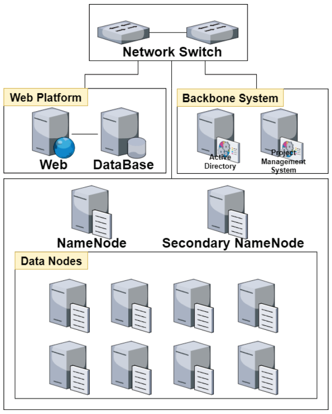

Figure 1 Physical configuration linking research data platform storage and web platform server.

configuration, and Hadoop cluster setup. Physical configuration refers to the setup of devices to be used for storage, whereas network configuration involves configuring the communication between each device. Finally, the Hadoop software and settings are used to complete the Hadoop infrastructure setup. For minimal construction of the physical configuration of the distributed storage infrastructure using Hadoop, one NameNode and three DataNodes are required. Physically, three servers (or computers), with one serving as both the NameNode and DataNode, are required. However, the recommended configuration calls for one NameNode, one secondary NameNode, and at least three DataNodes. It is also advisable to separate the roles into distinct physical devices when establishing the infrastructure.

In this study, we constructed the infrastructure shown in Figure 1, which comprises one NameNode, one secondary NameNode for NameNode fault tolerance, and eight DataNodes. The NameNode manages which DataNodes store the data and the actual data are divided into distributed blocks of configured block sizes, as illustrated in Figure 2. The Web Platform node is a node where web-based software for the user interface is executed. The Backbone System is an integrated system for user authentication and security. These blocks are replicated thrice across different data nodes. Even if one DataNode is excluded from the cluster because of a failure, the data can be retrieved using replicated information from other nodes.

The network configuration involved connecting each node using a single network switch. A 1 Gigabit network capacity was implemented, considering the utilization of large amounts of data and concurrent usage. In addition,

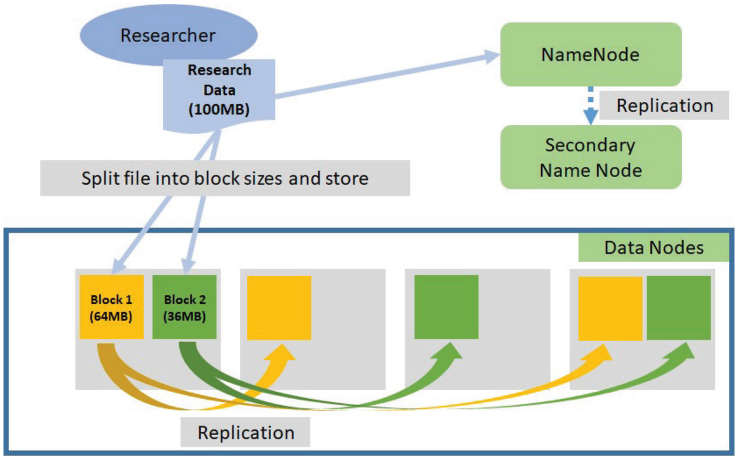

Figure 2 Example of block-level distributed storage in Hadoop distributed file system environment for research data.(The yellow and green rectangles explain the replication of blocks.)

SSH keys were mutually registered for intercommunication between the servers.

Finally, in the Hadoop cluster setup phase, the Hadoop configuration files, namely core-site.xml, hdfs-site.xml, yarn-site.xml, and mapred-site.xml, were modified according to the earlier setup by adjusting the network IPs and replication counts.

In this study, we configured a storage space of 1.8 PB (1920 TB) by equipping each of the eight data nodes with 12 SAS disks of 20 TB each. The replication was set to three, allowing for the utilization of 640 TB, which was one-third of the physical storage space. Furthermore, the advantage of the infrastructure built using distributed storage technology is its ease of scalability for future capacity expansion. Additional servers and configurations can be included to increase the capacity conveniently. Building such infrastructure using the Hadoop Distributed File System enables the management of various forms and formats of data in the field of food research. This, in turn, enables researchers to store and utilize data more efficiently and reliably.

3.2 Implementation of a Web-based User Interface and Access Control

Research data storage, as it is built within the infrastructure, lacks security and usability for direct use. To maximize the research data storage, we

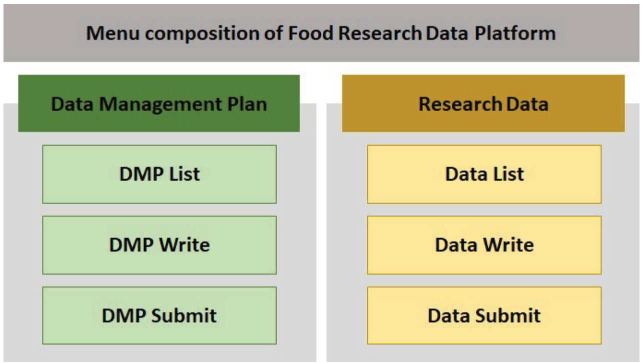

Figure 3 Menu structure diagram of the research data platform.

constructed a web-based data platform as a user interface. A web-based platform provides a means of managing access permissions for research data and allows for easy data uploading and sharing. The web-based data platform developed in this study revolves around DMPs and enables metadata management and data access control.

Figure 3 depicts the structure of the established research data platform. In the DMP menu, users can create metadata to understand the research data, including research methods, data types, formats, and production plans. In the Data menu, users can register the research data that are produced according to the DMP. Additionally, all of these functionalities include authorization controls to ensure that they can only be performed for the specific research projects in which the users are involved.

All platforms handle permission control through user logins. Various methods exist for processing logins, including in-house user databases and external integration. However, on this platform, we applied the method of querying user information and processing logins through an active directory, which is commonly used by enterprises for employee information management. Figures 4, 5, and 6 depict sequence diagrams. The rectangles and dotted lines represent Lifelines, while the vertical bars within rectangles signify Activations, illustrating interactions between Lifelines. As shown in Figure 4, when a user enters the login information, the platform verifies whether the user information in the active directory matches and then processes the login. If an in-house user database is used, this part can be modified to query the in-house information instead.

The platform is designed to allow the registration of research data with a DMP at its core. The DMP serves as a form of metadata for research

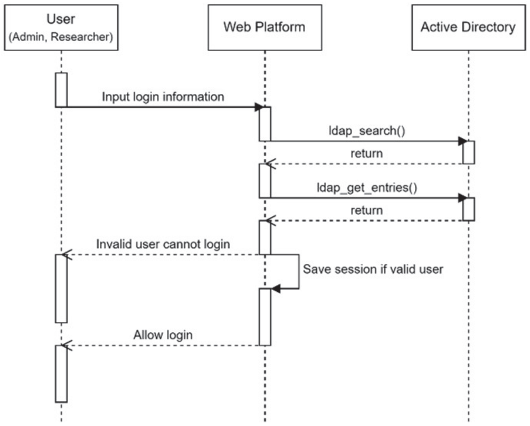

Figure 4 Procedure for controlling user permissions through login in the platform.

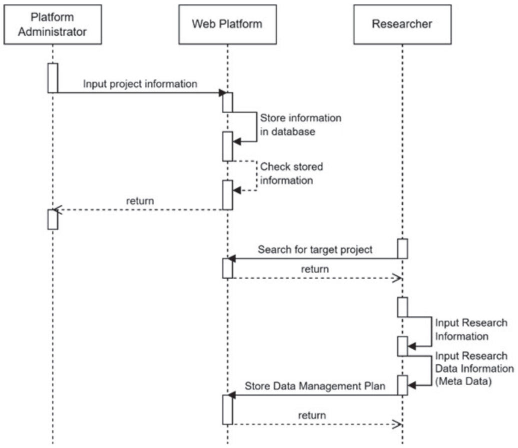
Platform Administrator
Web Platform
Researcher
Input project information
Store information in database
Check stored information
return
Search for target project
return
Input Research Information
Input Research Data Information (Meta Data)
Store Data Management Plan
return

Figure 5 Procedure for registering DMP in platform.

data. Researchers can define the format and structure of the research data by first registering a DMP and then eliminating the need for separate inputs when registering actual data. In addition, the registered metadata can be shared among researchers through information retrieval. Figure 5 depicts the procedure for registering a DMP for research data. Before researchers create a DMP, platform administrators input the project information into the web platform to grant permissions. The researchers then select the relevant project to which they have access and provide additional information to create a DMP.

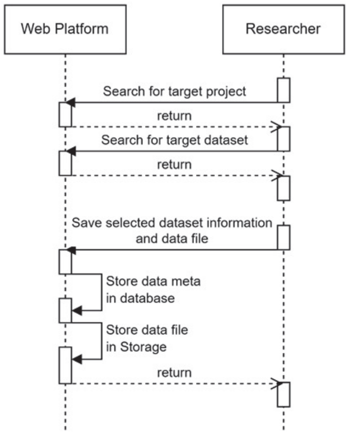
Web Platform
Researcher
Search for target project
return
Search for target dataset
return
Save selected dataset information and data file
Store data meta in database
Store data file in Storage
return

Figure 6 Procedure for registering research data in the platform.

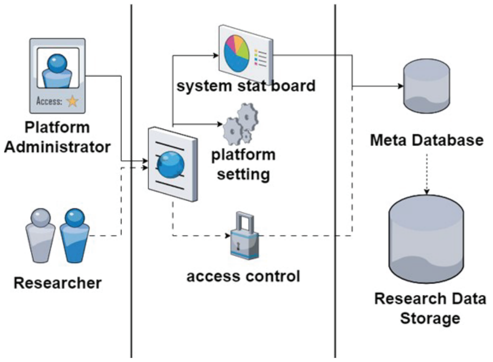

Figure 7 Diagram illustrating data storage based on the roles of administrators and researchers in a web-based data platform.

Once the DMP is completed, researchers can store the acquired data through research activities on the platform, similar to experiments. Figure 6 illustrates the procedure for registering research data in the platform. With the permissions obtained through logins, researchers can select the accessible DMP and register the actual research data by selecting the research data metadata that are registered in the DMP. Thereafter, the registered research data are stored in a distributed manner in the previously described storage system.

Figure 7 illustrates the usage and conceptualization of data storage for various roles using this platform. Researchers in the field of food science can easily access web-based platforms anywhere, at any time, using any device. They can store and manage data without being restricted by the format or structure.

4 Discussion

Food research encompasses a wide range of data types including experimental biological results, chemical analysis data, nutritional information, microbiological data, sensor data, images, and videos. Complete research data and data that are acquired for research purposes are valuable assets requiring secure management. Individual management by researchers or research teams has certain limitations. Furthermore, recent advancements in deep learning technology have made data sharing and collaborative usage inevitable. In the interdisciplinary field of food research, where various research areas converge, it is essential to establish a data sharing platform for the safe management and sharing of research data.

We have proposed a method to ensure the secure management of research data by building an infrastructure based on distributed storage using Hadoop. This method involves distributing and replicating research data in block-sized units to minimize the loss of research data due to system failure. In addition, we introduced a web-based data platform to enhance usability.

Although a web-based data platform is sufficient for storing, managing, and sharing data, its utility can be further enhanced when integrated with systems for data analysis. In future research, we plan to expand the research data platform to encompass the entire data lifecycle in food research by implementing methods for automatically collecting data generated by devices and web-sharing sites and incorporating commonly used web-based data analysis tools in food research.

5 Conclusions

In this study, we addressed the challenges of research data management and sharing in the field of food research by constructing a web-based platform and a Hadoop-based distributed storage system. The web-based platform was designed to accommodate various forms and formats of data in the food research domain, providing a secure storage environment and an easy means for researchers to share and utilize their data.

The research data platform offers researchers the ability to manage metadata relating to the research data, upload data, and control access permissions. It provides researchers with the freedom to access data securely from anywhere and anytime, thereby facilitating the efficient management and utilization of research data. Moreover, the Hadoop-based distributed storage system plays a crucial role in ensuring the secure preservation of data while minimizing data loss owing to system failures through distributed storage and replication.

We created a physical storage space of 1.8 PB using the Hadoop Distributed File System and built a platform with a web-based user interface. This platform is poised to drive innovation in data management and sharing within the field of food research, enabling researchers to perform their work more efficiently and contribute to the advancement of the field by utilizing data effectively. In the future, we plan to develop an extensible research data platform that encompasses the entire data lifecycle in food research through the integration of automated data collection and analysis tools. These efforts are anticipated to contribute to continuous innovation and development in the field of food research.

Acknowledgements

This research was supported by the Main Research Program (E0220700) of the Korea Food Research Institute (KFRI) funded by the Ministry of Science and ICT.

References

[1] Galanakis, C.M. (2020). The food systems in the era of the coronavirus (COVID-19) pandemic crisis. Foods, 9, 523.

[2] Jin, C., Bouzembrak, Y., Zhou, J., Liang, Q., Van Den Bulk, L.M., Gavai, A., Liu, N., Van Den Heuvel, L.J., Hoenderdaal, W., Marvin, H.J. (2020). Big Data in food safety- A review. Current Opinion in Food Science, 36, 24–32.

[3] National Center for Biotechnology Information. (n.d.). About NCBI. Retrieved from https://www.ncbi.nlm.nih.gov/home/about/ (accessed on 2023.10.5.).

[4] EMBL-EBI. (n.d.). About us. Retrieved from https://www.ebi.ac.uk/about (accessed on 2023.10.5.).

[5] DDBJ Center. (n.d.). About DDBJ Center. Retrieved from https://www.ddbj.nig.ac.jp/about/index-e.html (accessed on 2023.10.5.).

[6] National Genomics Data Center. (n.d.). About. Retrieved from https://ngdc.cncb.ac.cn/about (accessed on 2023.10.5.).

[7] kobic. (n.d.). About Us |Introduction. Retrieved from ttps://www.kobic.re.kr/kobic/intro/overview (accessed on 2023.10.5.).

[8] Foundation. A.S. (n.d.). Hadoop. Retrieved from https://hadoop.apache.org/ (accessed on 2023.10.5.).

[9] Ji, Q. (2021). A Novel Mass Meteorological Data Storage System Based on Hadoop Ecosystem. Fresenius Environmental Bulletin, 30(7), 5332–5339.

[10] Wu, J., Xiong, J., Dai, H., Wang, Y., Xu, C. (2022). MIX-RS: A multi-indexing system based on HDFS for remote sensing data storage. Tsinghua Science and Technology, 27(6), 881–893.

[11] Chawla, T., Singh, G., Pilli, E.S. (2021). MuSe: a multi-level storage scheme for big RDF data using MapReduce. Journal of Big Data, 8(1), 1–26.

[12] Sisodia, A., Jindal, R. (2022). An effective model for healthcare to process chronic kidney disease using big data processing. Journal of Ambient Intelligence and Humanized Computing, 1–17.

[13] Y. Chen, D. Li, L. Yan, Z. Ma. (2022). Two-Stage Detection of Semantic Redundancies in RDF Data. Journal of Web Engineering, 21(8), 2313–2337. doi: 10.13052/jwe1540-9589.2184.

[14] Chen, T., Ma, J., Liu, Y., Chen, Z., Xiao, N., Lu, Y., Fu, Y., Yang, C., Li, M., Wu, S. (2022). iProX in 2021: connecting proteomics data sharing with big data. Nucleic Acids Research, 50, D1522–D1527.

[15] Ferraro Petrillo, U., Palini, F., Cattaneo, G., Giancarlo, R. (2021). FASTA/Q data compressors for MapReduce-Hadoop genomics: space and time savings made easy. BMC Bioinformatics, 22(1), 1–21.

[16] Zu, C. (2021). Hadoop-Based Painting Resource Storage and Retrieval Platform Construction and Testing. Complexity, 2021, 1–11.

[17] Belov, V., Kosenkov, A.N., Nikulchev, E. (2021). Experimental characteristics study of data storage formats for data marts development within data lakes. Applied Sciences, 11(19), 8651.

[18] Armstrong, E.M., Bourassa, M.A., Cram, T.A., DeBellis, M., Elya, J., Greguska III, F.R., Huang, T., Jacob, J.C., Ji, Z., Jiang, Y. (2019). An Integrated Data Analytics Platform. Frontiers in Marine Science, 6, 354.

[19] Han, X., Shen, H., Hu, H., Gao, J. (2022). Open Innovation Web-Based Platform for Evaluation of Water Quality Based on Big Data Analysis. Sustainability, 14(22), 8811.

[20] Bossi, G., Schenato, L., Marcato, G. (2023). Web-Based Platforms for Landslide Risk Mitigation: The State of the Art. Water, 15(4), 1632.

[21] David, F.P., Litovchenko, M., Deplancke, B., Gardeux, V. (2020). ASAP 2020 update: an open, scalable and interactive web-based portal for (single-cell) omics analyses. Nucleic Acids Research, 48, W403–W414.

[22] Li, H., Shi, M., Ren, K., Zhang, L., Ye, W., Zhang, W., Cheng, Y., Xia, X.-Q. (2023). Visual Omics: a web-based platform for omics data analysis and visualization with rich graph-tuning capabilities. Bioinformatics, 39, btac777.

[23] Zhou, G., Ewald, J., Xia, J. (2021). OmicsAnalyst: a comprehensive web-based platform for visual analytics of multi-omics data. Nucleic Acids Research, 49, W476–W482.

[24] Zhou, G., Pang, Z., Lu, Y., Ewald, J., Xia, J. (2022). OmicsNet 2.0: a web-based platform for multi-omics integration and network visual analytics. Nucleic Acids Research, 50, W527-W533.

[25] Asif, M., Abbas, S., Khan, M. A., Fatima, A., Khan, M. A., and Lee, S. W. (2022). MapReduce based intelligent model for intrusion detection using machine learning technique. Journal of King Saud University-Computer and Information Sciences, 34(10), 9723-9731.

[26] Xiao, B., Yang, Z., Qiu, X., Xiao, J., Wang, G., Zeng, W., ... and Chen, W. (2021). PAM-DenseNet: A deep convolutional neural network for computer-aided COVID-19 diagnosis. IEEE Transactions on Cybernetics, 52(11), 12163–12174.

[27] Pavlova, M., Terhljan, N., Chung, A. G., Zhao, A., Surana, S., Aboutalebi, H., ... and Wong, A. (2022). Covid-net cxr-2: An enhanced deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Frontiers in Medicine, 9, 861680.

[28] Shinde, P. P., Desai, V. P., Katkar, S. V., Oza, K. S., Kamat, R. K., and Thakar, C. M. (2022). Big data analytics for mask prominence in COVID pandemic. Materials Today: Proceedings, 51, 2471–2475.

[29] Bawankule, K. L., Dewang, R. K., and Singh, A. K. (2022). Historical data based approach to mitigate stragglers from the Reduce phase of MapReduce in a heterogeneous Hadoop cluster. Cluster Computing, 25(5), 3193–3211.

[30] Amankwah-Amoah, J., Khan, Z., Wood, G., Knight, G. (2021). COVID-19 and digitalization: The great acceleration. Journal of Business Research, 136, 602–611.

Biographies

Gi-taek An received his B.Sc. in Computer Science from Namseoul University in 2011 and M.Sc. in Computer Science from Jeonbuk National University. Currently, he is a Senior Engineer at the Korea Food Research Institute and a Ph.D. student at Jeonbuk National University. His research areas include information retrieval, artificial intelligence, and data platforms.

Seyoung Oh received his B.Sc. in Computer Science from Hanbat University in 2010 and M.Sc. in Industrial System Engineering from Chungnam National University in 2018. Currently, he is an Engineer at the Korea Food Research Institute. His research area includes information systems and data platforms.

Eunhye Kim received her B.Sc. in Electronic Engineering from Jeonbuk National University in 2012. Currently, she is an Engineer at the Korea Food Research Institute. Her research area includes Information Systems.

Jung-min Park received her B.Sc. in Biology from Ewha Womans University in 1991. She received her M.Sc. and Ph.D. degrees in Economics from Hannam University in 2001 and 2005, respectively. Currently, she is Intelligent Policy Team Leader and Principal Researcher at the Korea Food Research Institute. Her research areas include technology innovation and research data.