Data Framework Design of EDISON 2.0 Digital Platform for Convergence Research

Sunggeun Han[1], Jaegwang Lee[1], Inho Jeon[1], Jeongcheol Lee[1] and Hoon Choi[2*]
[1] Korea Institute of Science and Technology Information
Daejeon, South Korea
[e-mail: {sghan, leejg, inojeon, jclee}@kisti.re.kr]
[2] Chungnam National University
Daejeon, South Korea
[e-mail: hc@cnu.ac.kr]
*Corresponding author: Hoon Choi

Received February 26, 2023; revised July 13, 2023; accepted August 17, 2023; published August 31, 2023

Abstract

With improving computing performance, various digital platforms are being developed to enable easily utilization of high-performance computing environments. EDISON 1.0 is an online simulation platform widely used in computational science and engineering education. As the research paradigm changes, the demand for developing the EDISON 1.0 platform centered on simulation into the EDISON 2.0 platform centered on data and artificial intelligence is growing. Herein, a data framework, a core module for data-centric research on EDISON 2.0 digital platform, is proposed. The proposed data framework provides the following three functions. First, it provides a data repository suitable for the data lifecycle to increase research reproducibility. Second, it provides a new data model that can integrate, manage, search, and utilize heterogeneous data to support a data-driven interdisciplinary convergence research environment. Finally, it provides an exploratory data analysis (EDA) service and data enrichment using an AI model, both developed to strengthen data reliability and maximize the efficiency and effectiveness of research endeavors. Using the EDISON 2.0 data framework, researchers can conduct interdisciplinary convergence research using heterogeneous data and easily perform data pre-processing through the web-based UI. Further, it presents the opportunity to leverage the derived data obtained through AI technology to gain insights and create new research topics.

Keywords: Computational Science, Convergence Research, Data-driven Research, Data Framework, Digital Platform

A preliminary version of this paper was presented at ICONI 2022, and was selected as an outstanding paper. This research was supported by the EDISON Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. NRF-2022M3C1A6090416). Hoon Choi was supported by Regional Innovation Strategy (RIS) through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (MOE) (2021RIS-004).

http://doi.org/10.3837/tiis.2023.08.019 **ISSN : 1976-7277**

1. Introduction

Computational science uses a method of interpreting mathematical models through computer calculations rather than conventional theoretical and experimental methods to gain an understanding of a research topic [1]. It involves the development of models and simulations that require enormous amounts of computation [2]. Recently, with the rapid development of computer performance and cyberinfrastructure, simulations at a previously impossible level and scale are possible through large-scale parallel calculations. Further, a new field of R&D has emerged through the accumulation and utilization of data, and the paradigm of scientific research exploration has shifted from a trial-and-error approach through experiments, theories, and simulations to data-driven research. The ever-increasing importance of data sometimes leads to a "data flood" [3,4]. Additionally, numerous research and publishing communities demand data sharing and public access to data. Increasingly more efforts are being devoted to sharing large amounts of data, techniques, and tools generated by research through open science and open access to facilitate collaborative and interdisciplinary research [5,6].

However, the biggest bottleneck to this is the lack of time and resources required to easily collect, transform, filter, and upload data. Researchers demand an open science platform where they can use existing data and submit new content without requiring technical expertise. Consequently, various digital platforms have emerged to render ease in sharing, leveraging, and collaborating on research data. Researchers across the world can easily access data, tools, and simulation resources via the digital platform. Additionally, they manage their workflow; this ease in transferring or replicating data in a computing lab environment allows the researchers to focus on scientific research [7,8]. Recently, with the rapid development of artificial intelligence (AI) technology, attempts to combine simulation and AI technologies have increased in the field of computational science. However, it is significantly difficult for domain researchers to use; for example, acquiring a programming language to AI technology or building an AI algorithm execution environment. These barriers are also opportunities to easily integrate and use AI technologies in research by leveraging digital platforms to uncover new insights and research challenges [9,10].

Education-research Integration through Simulation On the Net (EDISON) is an online simulation platform for teaching and researching in science and engineering professional applications; it is widely used by professors, students, and researchers in education, including simulation software and educational content. EDISON was developed as a simulation-oriented computational science platform and contributed to the formation and activation of the computational science community for programs or software that only a small number of scientists use for research [11,12]. Currently, EDISON is being further developed in response to the demand for a change from a simulation-centered education platform to a data- and AI-centered research platform. Therefore, in this study, the first version of EDISON is named EDISON 1.0, and the new EDISON is named EDISON 2.0.

During to the operation of EDISON 1.0 in the past decade, 900 types of simulation software, 800 types of educational content, and considerable amount of data were accumulated to form a research community in various fields. Consequently, the need for interdisciplinary convergence research on accumulated data has increased [13,14]. Therefore, in this study, three tasks were set and solutions to transform simulation-oriented EDISON 1.0 into data-oriented EDISON 2.0 were proposed, as follows.

- EDISON 2.0 should be able to support the entire end-to-end research process; this is related to research reproducibility [15,16]. Typically, a computer science researcher has a research lifecycle in each discipline: data collection, analysis, execution, result

comparison, and publication [17]. The results of each stage of the research pipeline must be stored and managed to express and execute the entire research process.

- For interdisciplinary convergence research, data in various fields must be easily managed. Further, retrieving, merging, transforming, and processing disparate data should be easy. This requires the provision of data models and representations in an integrated manner. In this study, integrated metadata were provided for this purpose.
- It should be possible to reproduce new data by applying the latest AI technology to the data, and an environment to analyze large-scale data produced through this should be provided [18]. By applying AI technology based on researchers' data to generate new derived data, insights can be gained, and opportunities for new research challenges can be created.

The remainder of this paper is organized as follows. Section 2 reviews the computational science platforms and related data platforms. A general overview of EDISON 2.0 and its data framework are described in Section 3. Section 4 describes a data repository that supports the data lifecycle, and Section 5 addresses the heterogeneous data integration environment for interdisciplinary convergence research. Section 6 addresses the exploratory data analysis (EDA) service and data enrichment using AI model. Finally, Section 7 concludes the study and discusses future work.

2. Related Work

A digital platform is an online platform that enables the exchange of goods, services, and information between different users, as well as a digital infrastructure that enables interaction and collaboration between users [19,20]. Examples of digital platforms include e-commerce platforms such as Amazon, social media platforms such as Facebook, and sharing economy platforms such as Airbnb. These platforms are increasing user convenience and transaction efficiency while creating new business opportunities [21].

Recently, with the rapid development of AI technology, a new digital platform called an AI platform has emerged. AI platforms are digital platforms that provide services based on AI. It provides machine learning, deep learning, and various data analysis tools to help users develop, learn, test, and deploy predictive models. Major companies such as Google, Amazon, and Microsoft, with their powerful cloud infrastructure and extensive AI research and development capabilities, are leading the development of AI platforms. Amazon SageMaker is a fully managed machine learning service on AWS that makes developing, training, and deploying machine learning models easy [22]. It also manages all the infrastructure needed to train and deploy models, freeing developers to focus on their own work. Microsoft Azure AI provides cloud-based AI services [23]. It includes services such as machine learning, natural language processing, computer vision, and speech recognition, and it supports the development, training, and deployment of machine learning models. Google AI Platform provides AI developers with the tools they need to create and train models [24]. It can be used in conjunction with Google's powerful data analytics and machine learning services, and supports leading machine learning libraries such as TensorFlow, scikit-learn, and XGBoost. These platforms automate or create new services that previously required human intervention [9].

Computational science platforms are digital platforms that support research in the field of computational science, which requires a variety of tools and services not typically supported by AI platforms. While AI platforms are primarily used by data scientists to perform complex data analysis and modeling, computational science platforms provide the tools and services

needed to solve a considerably broader range of scientific problems. With the increasing use of AI technologies in computational science, efforts are ongoing to integrate AI tools into existing computational science platforms. This integration has the dual benefit of providing access to AI tools and leveraging the extensive resources offered by computational science platforms. In this paper, we present computational science platforms from a data perspective, and in the following sections, we review the major computational science platforms and take a close look at how they support data management capabilities and how they are leveraging AI technologies.

2.1 HUBzero, PURR, and MyGeoHub

HUBzero is a computational science and engineering platform software developed by Purdue University and NCSA, with support from the US National Science Foundation (NSF). It is actively used as a platform for sharing programs and data in 60 research communities, including physics, biology, healthcare, natural science, pharmacy, and climatology [25,26]. Users can easily run simulation and modeling tools online using HUBzero, and the visualized results can be viewed directly through a web browser. Furthermore, it provides social networking capabilities for researchers to collaborate and the ability to store, search, and share data. HUBzero is a web-based science and engineering collaboration platform that provides various features: simulation-modeling-analysis tool hosting, data publishing, resource sharing, community organization and collaboration, data analysis, and machine-learning tools (Jupyter Notebook and RStudio).

The Purdue University Research Repository (PURR) was developed by Purdue University as a platform for sharing and managing research data within universities and customized HUBzero [27]. Essentially, PURR is a custom instance of HUBzero, which supports scientific discovery, learning, and collaboration. Through the PURR platform, researchers can disseminate data for public access and discovery. The HUBzero platform has been developed without considering metadata or preservation. Therefore, custom metadata implementation should be used as a trusted digital repository and has become one of the development goals of PURR. PURR has developed custom metadata schemas such as metadata encoding and transmission standard (METS), Dublin core initiative metadata (dcterms), metadata object description schema (MODS), and preservation metadata implementation strategies (PREMIS).

MyGeoHub extends HUBzero's file-publishing capabilities to publish DOI assignments and services for a set of files, including metadata for all files [28]. Data management provides an iData data management infrastructure based on the iRODS data management server, metadata extraction from highly structured spatial files, keyword and spatial range searches, and spatial file visualization capabilities [29]. MyGeoHub uses MultiSpec to create complex data views or perform significant data processing, analysis, or transformation and supports a variety of image formats. In addition, it provides several visualization toolkits, including a sandbox environment that enables users to overlay geospatial data on world maps, without any programming skills, using the GeoBuilder toolkit. A REST API is available in MyGeoHub, which allows third-party applications to access the iData file management system and conduct all file management operations, such as uploading, downloading, listing, renaming, and metadata management of files. It supports the interoperability between gateways that can use tools and tools from other gateways. MyGeoHub proposed a sustainability model, provided HUBzero's scientific data management, certification and high-performance computing (HPC) resource management capabilities, and delivered a comprehensive tool development environment to users, allowing them to easily introduce scientific code online.

2.2 Materials Project and MPContribs

The Materials Project, a component of the U.S. Materials Genome Initiative, leverages HPC to compute the structural, thermodynamic, electronic, and mechanical properties of more than 60,000 inorganic compounds using high-throughput ab-initio computations [30,31]. The computation outcomes and analysis utilities are shared with the public via contemporary web and application interfaces. These outcomes and tools may facilitate the acceleration of discovering, designing, and producing advanced materials for applications such as batteries, photovoltaics, and semiconductors. However, as the number of users increases beyond 10,000, community-driven data submissions are required, which should expand the coverage and improve the integrity and quality of diverse datasets.

MPContribs is a computing software infrastructure that integrates and organizes user-supplied simulated or measured material data [32]. MPContribs integrates standard data and community-provided datasets to build user communities, enables integrated search across datasets, and provides processed and interpretable data. Data processing is integrated and configured with the existing collaborative graph platform, whereas maintenance and formatting are controlled by the user. It provides data retrieval mechanism through a REST API and computes aggregates of submitted datasets for use in integrated analytics.

2.3 AiiDA and Materials Cloud

A key goal of AiiDA is the complete reproducibility of the computational and resulting data obtained through a tight coupling of storage and workflow automation. This allows researchers to accelerate the computational science process and eliminate considerable details and techniques of error-prone simulations while ultimately providing an open-access model for computation [33]. AiiDA proposed the ADES model. The automation model allows the daemon to operate in the background and handle interactions with the HPC cluster. In data models, calculations and data are represented as graph nodes and heterogeneous data are accommodated using entity-attribute-value (EAV) tables. The environment model supports visualization, text parsing, and scientific data processing through Python libraries and provides AiiDA plugins, Verdi command utilities, scientific workflow, and query tools [34]. The sharing model provides the ability to easily share tools and results such as data, code, and workflows.

The aim of Materials Cloud is to establish an ecosystem that supports researchers throughout the lifespan of a scientific project while promoting reproducible and equitable research findings [35]. AiiDA works similarly to Git's scientific workflow tracking, whereas Materials Cloud works similarly to the GitHub platform for sharing, discovering, and visualizing anything tracked by AiiDA. Materials Cloud establishes a global schema for accumulating all submissions into a single database and adopts a model known as "storage repository," which is analogous to GitHub, giving each submission its designated space. It provides a comprehensive platform for open science that is accessible to anyone free of charge, offering five sections: LEARN, WORK, DISCOVER, EXPLORE, and ARCHIVE.

2.4 EDISON 1.0 and EDISON-SDR

EDISON 1.0 is a platform developed as an advanced science education hub development project (Education-to-industry integration through simulation on the open platform and Net, EDISON) [11]. Based on the cyberinfrastructure composed of large-scale computing and network resources, professors, students, researchers, and industrial workers in the field of computational science and engineering share and run simulation programs and contents such

that anyone can easily conduct education and research. The EDISON platform supports software registration and execution, data sharing, and community collaboration functions in seven fields (computational thermal fluids, nanophysics, computational chemistry, structural mechanics, computational design, computational medicine, and urban environments). The tools provided by EDISON are as follows scientific app registration and management, user authentication (security), basic portal creation, cyberinfrastructure integration workflow execution, and simulation result visualization tools. **Fig. 1** illustrates the overall architecture of EDISON 1.0.

EDISON-SDR (Science Data Repository) extends the EDISON 1.0 platform to provide services that allow students to learn data-driven research methodologies and easily publish, share, search, and analyze computer simulation data [36]. EDISON-SDR provides an automated preprocessing framework to solve the problems of complexity, diversity, reliability, connectivity, and heterogeneity of computational science data, data expression methods by data type, data quality management, data group management, and metadata expression technology.

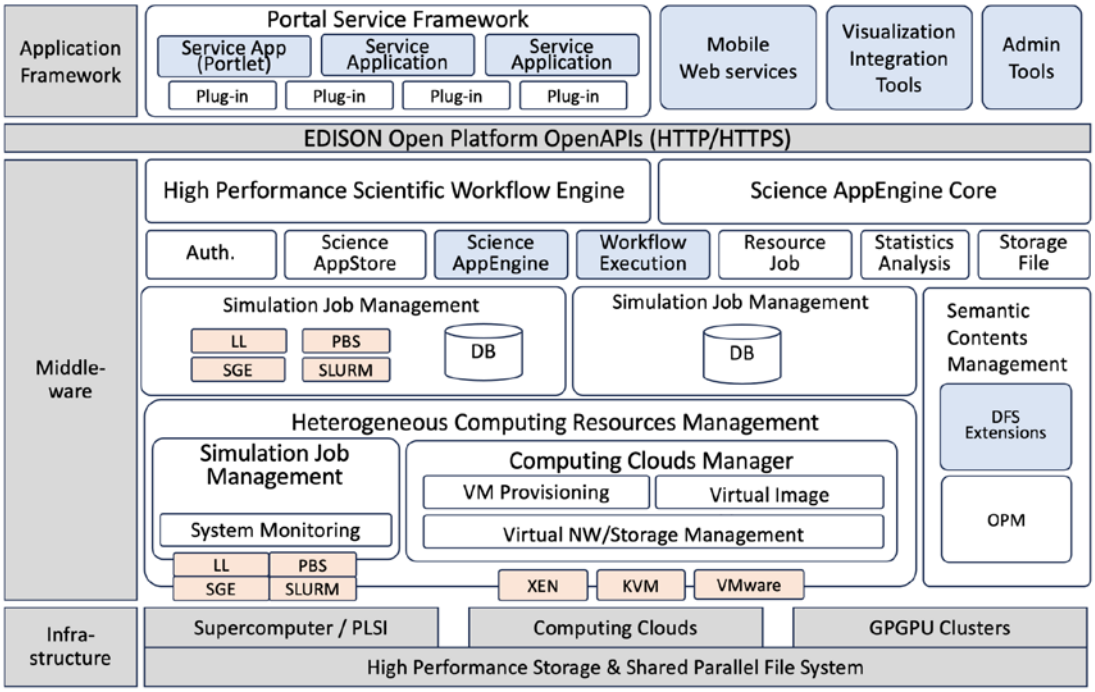

Fig. 1. Overall system architecture of EDISON 1.0

2.5 Evaluating Data Capabilities Across Platforms

In this section, we evaluate and compare data capabilities of the aforementioned computational science platforms. The following data features are compared:

- Does it provide metadata for data integration? Does it provide custom metadata or schema authoring tools?
- Does it support a data lifecycle for research reproducibility? Does it have a separate data repository?
- Does it provide a data analysis environment?
- Does it provide data analysis capabilities utilizing AI technology?

These features were derived while designing the data framework to expand the EDISON 1.0 platform into a platform for data-centered convergence research. **Table 1** shows a comparison of data capabilities across platforms.

The HUBzero platform provides metadata for data storage, search, and sharing, alongside the Simulation workflow, modeling tools, and AI technology such as Jupyter notebooks and RStudio. However, it does not offer custom metadata tools, a data-driven lifecycle, or a data analysis environment. The PURR platform supports customizable metadata and workflow for data publishing, but it does not provide separate data lifecycle storage, data analysis environment, or AI utilization. MyGeoHub facilitates metadata extraction for spatial files, and the iData system supports large file processing and analysis, with Jupyter notebook for AI use; however, neither offer data lifecycle features.

The Materials project platform offers metadata, APIs for inorganic-compound integration, and tools for calculating and visualizing materials properties, and it provides data for training machine learning algorithms, but it lacks custom schema tools or data lifecycle support. The MPContribs platform integrates community-contributed datasets, supports data publishing and management workflows, and provides data analysis in pandas data-frame format, but it does not offer data lifecycle functionality or AI-related features.

The AiiDA platform offers metadata for database calculation results, custom-data-type support, storage, workflow automation, Python libraries for data analysis, and AI utilization through AiiDAlab's Jupyter notebooks. The Materials Project platform uses a global schema for unified storage, provides Dublin Core metadata, a "repository" model for reproducibility, and Python libraries for both data analysis and machine learning for AI use.

EDISON 1.0 offers metadata for data storage, sharing, and search; it supports operation lifecycle via processing workflow and allows user-installed data analysis modules through its Workbench plugin, but it lacks separate data lifecycle functions, storage, and AI capabilities. EDISON-SDR provides metadata for data representation and management, supports data processing lifecycle via curation, and offers Jupyter Notebook for AI applications, but it does not provide separate storage or data analysis features.

EDISON 2.0 offers unified metadata with customizable schemas, workflow, data curation, lifecycle-based storage, and data preprocessing tools for AI models, and it can process large data via EDISON clusters. It facilitates AI-technology integration with a framework-executable environment and built-in AI algorithms for data analysis.

Table 1. Comparison of data capabilities

Platform	Metadata & Schema for Integration	Data Lifecycle for Research Reproducibility	Data Analysis Support	AI-based Technology Provision
HUBzero	- Integration Metadata (for data storage, search, and sharing)	- Simulation workflow	- Modeling & Simulation tool	- Jupyter Notebook - RStudio - Scientific library
PURR	- DC metadata - Custom metadata	- Data publication workflow	- N/A	- N/A
MyGeoHub	- DC metadata	- N/A	- Geospatial data analysis tool	- Jupyter notebook
Materials Project	- Pre-defined metadata	- N/A	- Materials analysis tool (Phase diagram, Crystal Toolkit, Reaction Calculator)	- ML library
MPContribs	- Standardized metadata - Community metadata	- Data publication workflow	- Python library	- N/A

AiiDA	- Custom metadata	- Simulation workflow - Repository automation	- Python library	- Python library - Jupyter notebook
Materials Cloud	- DC metadata - Global schema	- Research lifecycle - Repository lifecycle	- Python library	- ML library
EDISON 1.0	- Integration Metadata (for data storage, search, and sharing)	- Simulation workflow	- Workbench plugin	- N/A
EDISON-SDR	- Integration Metadata (for data representation and management)	- Data curation	- N/A	- Jupyter notebook
EDISON 2.0 Data Framework	- Custom metadata - Custom schema - Schema authoring tool	- Simulation workflow - Data lifecycle repository	- Exploratory Data Analysis (EDA) service	- Built-in AI-based data analysis algorithm - Use AI model by AI Framework

3. EDISON 2.0 Data Framework Overview

Different observational methods have been used in different fields of scientific research. **Fig. 2** shows the data flow of scientific research using the different approaches. Traditionally, experimental scientists produce data by observing subjects with the naked eye, using high-tech equipment, or by observing samples through experiments. Theoretical scientists create mathematical models of research objects, implement them in simulations, observe the objects, and generate data. A data scientist analyzes the observations (or data) collected through statistical methods or machine learning. **Fig. 2** indicates that scientists from various fields can gather in one place because data is the center. This can be a good example showing that data play an important role in interdisciplinary research.

In the field of big data, numerous advances have been made in periodically processing incoming data, such as processing sensor data or logging data; the most important function of data platform is the storage function that stores and retrieves the data itself. To analyze the data, the researchers downloaded the data locally, ran a local analysis program, and published the best results in a journal. Therefore, no additional functionality, and the concept of accessing or sharing problems that arose during the experiments or analyses were not required. However, in the field of computational science, because the size of the data is extremely large, calculation optimization or parallel efficiency is important, and storing all repeated data indefinitely is impossible. Deciding the data that can be discarded and that should be retained is a critical issue. Therefore, when handling data, the different characteristics of general and computational science data must be known. Computational science data platforms should manage data regarding whether any part of a running RUN fails during a simulation or diverges rather than converges in the middle of a computation. Moreover, everything used in research scenarios, such as research results, workflow, and jupyter notebook contents, should be managed, and functions to search, share, and utilize should be provided.

As described in the previous section, EDISON 1.0 was initially developed as a simulation-based educational platform, which transitioned to a research-focused EDISON 2.0 due to changing requirements. This new version not only expands upon the functionalities of its predecessor but also incorporates three main frameworks: simulation, data, and AI. **Table 2** details these enhancements and additions in EDISON 2.0 based on the primary features requested, and **Fig. 3** depicts the overall system architecture of this upgraded platform.

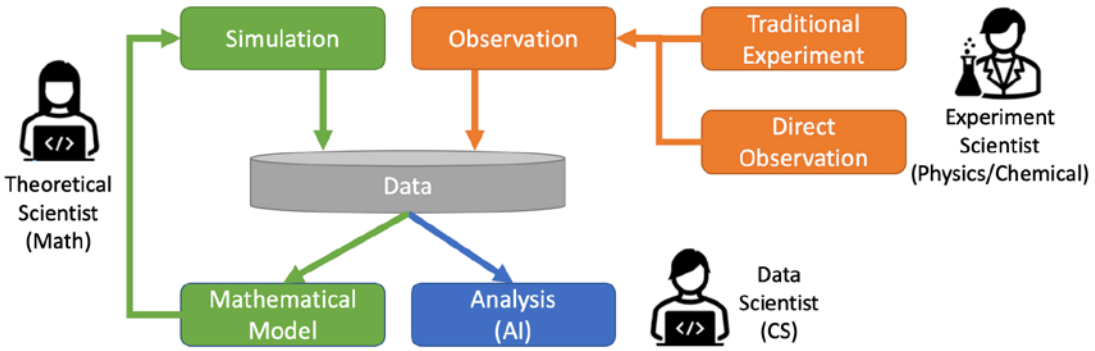

Fig. 2. Data flow in scientific data research

Table 2. EDISON 1.0 and 2.0 functionalities

Framework	Required Functionality	EDISON 1.0	EDISON 2.0	Benefits
Simulation Framework	- Computational science software development environment	- A public dedicated development server	- A per-user container-based development environment	- Troubleshoot user library installation - Troubleshoot machine dependencies
	- Customize user interface	- Workbench and workflow UI	- Workbench, workflow UI upgrades - Customizable UI: Terminal services, Custom App, JupyterLab, Desktop App, etc.	- A more selective and customizable UI
	- Computational science software execution environment	- VM-based HPC clusters for distributed processing of large jobs	- Cloud environment consisting of HPC clusters and container-based K8S clusters	- Efficiently manage resources and run large jobs
Data Framework	- A data repository supporting reproducibility	- Unified storage for data registration and sharing	- A repository with data lifecycle support	- Efficiently manage data storage and utilization, enable research reproducibility
	- Heterogeneous data integration environments	- Manage data registration based on pre-defined data types	- Schema-based heterogeneous data integration management - Support for standard schemas per research domain	- Heterogeneous data with custom schema support
	- Advanced data for AI models	- N/A	- Enrich data with AI models	- Enrich data by applying the latest user-developed AI techniques and models to data
	- Exploratory data analysis (EDA) environments	- N/A	- Data analysis service for data preprocessing - Handling large amounts of data through EDISON scheduler	- An easy data analysis environment - EDA tools for preparing input data for AI model - Efficient processing of large amounts of data
AI Framework	- Web-based AI model development environment	- Jupyter Notebook	- JupyterLab environment - Model registration service - Data linkage service	- AI models using shared data - Sharing AI models - Speed up AI model development

	- Web-based AI service development environment	- N/A	- AI service registration - Deploying AI packages	- Shared AI services can be directly utilized for a variety of studies
	- Cloud delivery	- Working with HPC clusters	- Cloud-based AI model training and services	- Data processing and AI training at scale

The simulation framework in EDISON 2.0 extends the HPC-cluster-based workbench and workflow to the cloud, providing customizable UIs. The data framework encourages research reproducibility through lifecycle-reflective storage organization and offers tools for exploratory data analysis and preprocessing, thereby reducing total data analysis time. The AI framework enhances Jupyter notebooks to JupyterLab, simplifying AI model registration, service registration, and package deployment.

At the heart of EDISON 2.0 is the integration among these frameworks, which facilitates data storage, sharing, simulation, and model training. The ability to import AI models within the data framework promotes data enrichment, leading to high-quality datasets utilizing cutting-edge AI models. The data framework also enables efficient parallel processing of extensive datasets through the EDISON scheduler offered by the simulation framework. This interconnected structure fosters data-centric integrated research, underscoring the pivotal role of data in facilitating interdisciplinary collaboration.

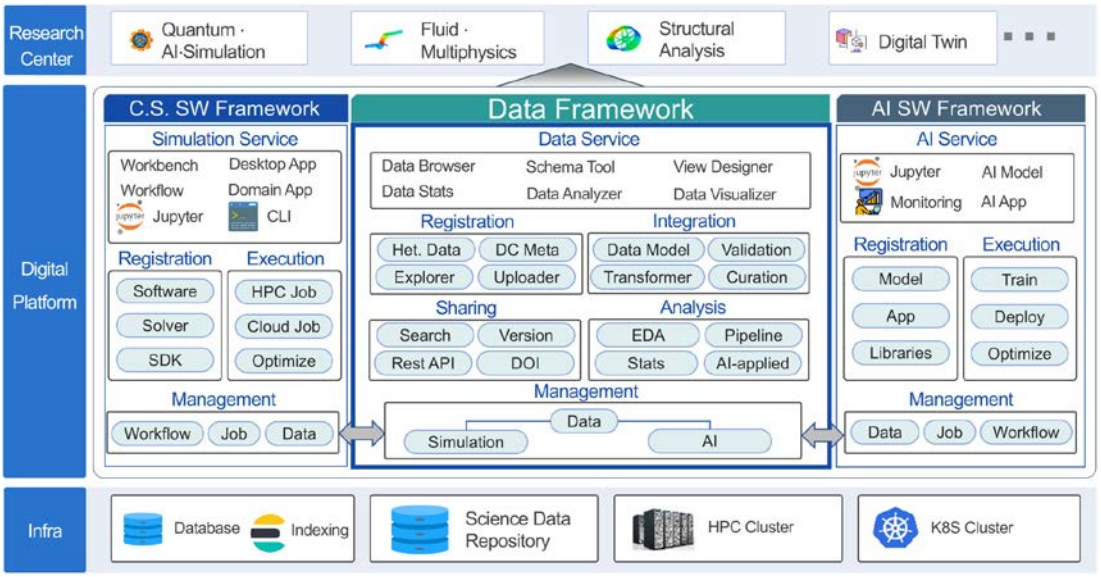

Fig. 3. Overall architecture of EDISON 2.0

4. Data Repository with Data Lifecycle

As mentioned above, data management in computational science requires a different approach than that required in general data management. The parameters used in the experiment may be more important than the storage and management of the entire dataset. Researchers have their research processes, lifecycles, and pipelines. Essentially, important information at each stage of the research must be tracked, stored, and managed. Furthermore, the input parameters are more important than the code files used in the experiment or data. Therefore, simplified code configuration management may be required instead of version control, which compares all

data files. The research lifecycle refers to the process of conducting research through project planning, execution, conclusion, publication, and deployment of research results. During the research lifecycle, data in the broadest sense are involved, including experimental, observational, acquired, and simulated data, and relevant information, artifacts, and original sources. The research lifecycle also includes published data, codes, and workflows to promote the reproducibility of published results. **Fig. 4** shows the USGS science data lifecycle model [37, 38].

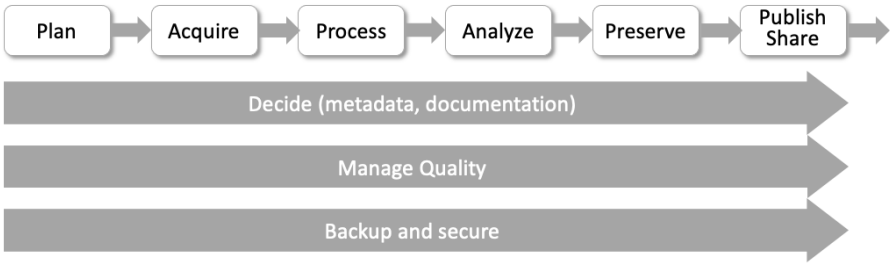

Fig. 4. USGS science data lifecycle model

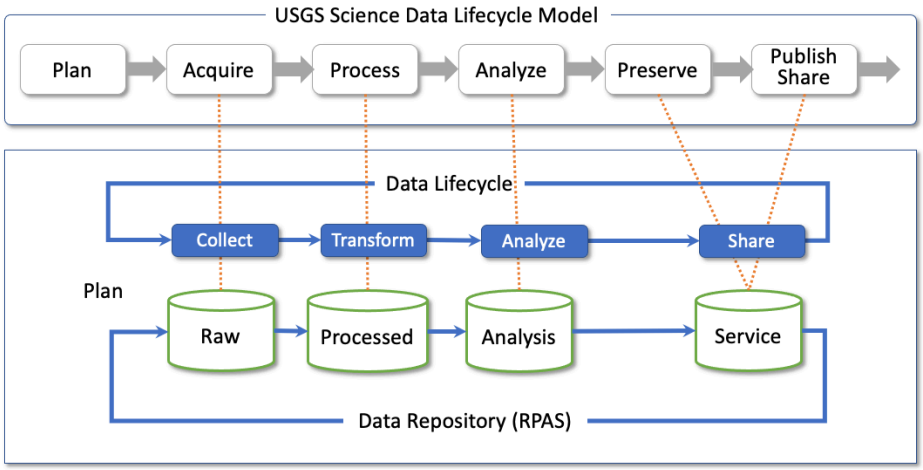

Fig. 5. EDISON 2.0 data repository

The data generated during the research lifecycle automatically creates a data lifecycle. As the importance of sharing data is emphasized, the data lifecycle is becoming an important factor. In the data lifecycle model, the data generated by the research lifecycle are planned, collected, processed, analyzed, preserved, published, and shared for reuse by others.

In addition to these activities, documenting workflow processes, providing metadata, and backing data should be performed continuously at all stages of the data lifecycle. **Fig. 5** shows the application of the EDISON 2.0 data repository to the USGS science data lifecycle model. In data-driven research, data are not fixed but undergo a process of changing into various forms through research or experimentation. Essentially, it follows the lifecycle of data collection, transformation, analysis, and sharing, respectively [39]. Raw data are converted into more valuable data through the lifecycle and serviced, and data that have reached the end of their lifespan are stored as an archive or derived through new data processing techniques and used as raw data. The EDISON 2.0 data repository applies these data attributes and includes the configuration management of the data accumulated at each stage. The states of the data in the data repository are as follows:

Raw data. These refer to the first data registered by the researcher or collected from the system, and refers to the original data that has not changed in its initial state.
Processed data. Although raw data are meaningful in their right, researchers typically process them for experiments or analysis software. The processed data can be generated by any number of algorithms or software used by researchers.
Analysis data. These refer to the data generated from analyzing the software used by the researcher. This data, created mid-research and easily discarded in the past, is now considered critical for AI-driven research.
Service data. Service data refer to the data published as the final research results. These data are important for the reproducibility of research. Additionally, they serve as an archive when service publications expire. Recently, service data have been converted into recycled data for interdisciplinary research and are being used as important data for convergence research, along with various heterogeneous data.

Data repositories designed around the research and data lifecycles enable researchers to capture the data generated at each stage of the research process. Further, they have the advantage of being able to understand the progress of research and realize the reproducibility of research through data monitoring and easy data management.

5. Heterogeneous Data Integration Environment

Typically, in data retrieval, the important factor is title matching; however, in computational science, it can be the information in the content inside the data. It is about how accurately hidden information can be found in computational science data. Particularly, when searching for heterogeneous computational data in various fields, the entire search target data must be generalized; however, generalizing data with different characteristics is challenging. One convenient approach to achieve this is if researchers can easily extract and view the data they want to see, such as by collecting specific variants from an entire dataset or classifying data by unique properties in metadata. A platform capable of grouping data in this manner, providing schemas suitable for queries of researchers in each field, and enabling easy collaboration among expert groups can endow researchers with the advantages of convenient searching and significantly high analysis efficiency.

EDISON 1.0 provides software from seven specialized centers and heterogeneous data by field. For interdisciplinary convergence research on heterogeneous data, a new data model that can integrate and manage the data is required. To this end, EDISON 2.0 designed new integrated metadata by defining common metadata, characteristic metadata, and schema. **Fig. 6** shows the integrated metadata proposed in this study.

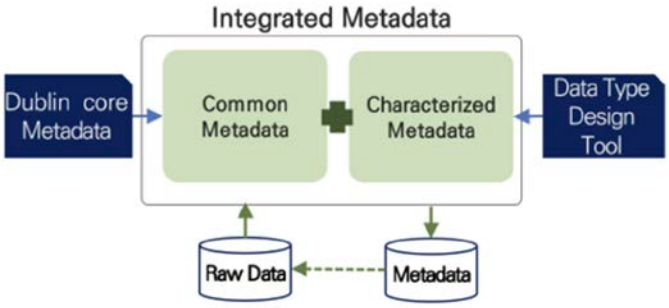

Fig. 6. Integrated metadata

Common metadata represent basic information for representing all the collected data. For example, data name, data description, creation date, and author are included. Common metadata consist of basic and additional metadata. Basic metadata consist of information that should be provided when registering data based on the Dublin core metadata [40]. Additional metadata consist of various elements required for open data and open access, such as proprietary information, related information, file information, and intelligence information. Particularly, intelligent information (or smart information) is data extracted through artificial intelligence technology and can be used as important information for interdisciplinary convergence research. **Table 3** lists common metadata.

Characteristic metadata represent information that expresses different characteristics that are considered important in each research field. As each research field has different types of data or characteristics of interest, the "Data Type" is defined and used. The data types are used as follows. (1) It serves as a data schema and is used for standardized data entry restrictions and data validation. (2) A visualization screen can be created using the "Data Type Design Tool." **Fig. 7** shows a visualization screen using the data-type definition and data-type design tool for OQMD data [41].

Table 3. Common metadata

Information	Metadata name	Description	Metadata name	Description	Metadata name	Description
Basic	**Title**	dataset name	**Description**	description	**Keyword**	keyword
	Language	Language	**Date**	creation date	**Category**	dataset category
Proprietary	**Creator**	data creator	**Contributors**	contributors	**Provider**	provider
	License	license	**Geospatial Coverage**	collection area	**Visibility**	disclosure
	Publish date	publish date				
Relational	**Source**	related data	**Source ID**	related data id	**DOI**	digital object identifier
	RID	national researcher ID	**SID**	science and technology standard classification ID	**NTISID**	NTIS ID
	DMP	data management plan	**JID**	Journal ID	**EID**	experiment ID
File	**Data type**	data type	**File format**	file format	**Dataset Info**	file configuration
	RPAS	file status	**Version**	file version	**Q score**	data qualification
Intelligence	**Auto ext**	auto-extract	**Auto cls**	auto-classification	**Auto rel**	auto-relation
	Auto rec	auto-recommend				

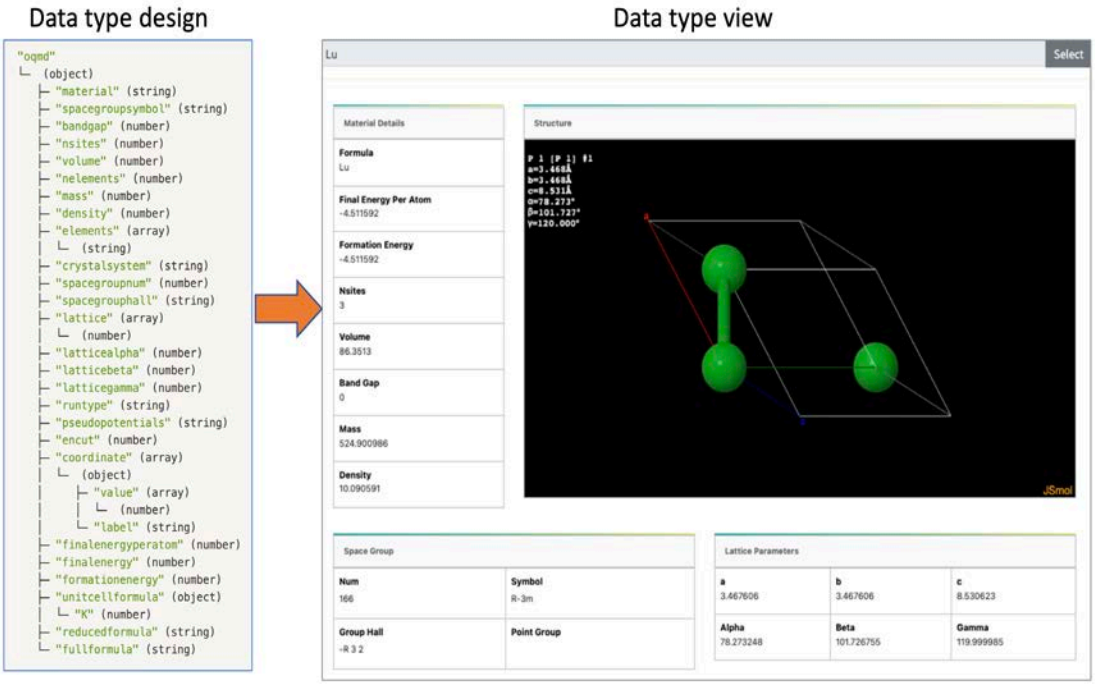

Fig. 7. OQMD data type definition and view

EDISON 2.0 configures a heterogeneous data integration environment based on integrated metadata and uses it in data standardization, data verification, data processing automation, integrated search, and data quality management modules. **Fig. 8** shows the heterogeneous data integration environment of EDISON 2.0.

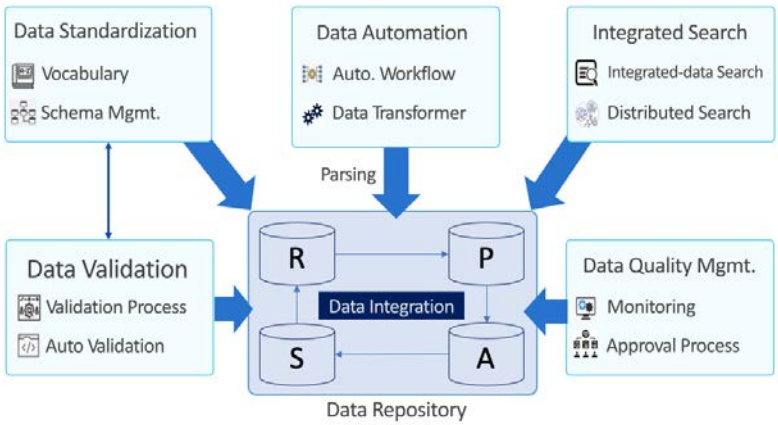

Fig. 8. Data integration environment

In a heterogeneous data-integration environment, researchers can conduct data-driven research by selecting well-structured data types suitable for their research. **Fig. 9** shows the DFT simulation process based on the XSF data type in EDISON 2.0. The DFT simulation work can be divided into three modules: structure builder, simulator, and analyzer. Considering these modules as one unit, each module can be added based on the purpose of the simulation, and data suitable for the researcher can be produced by linking the inputs and outputs.

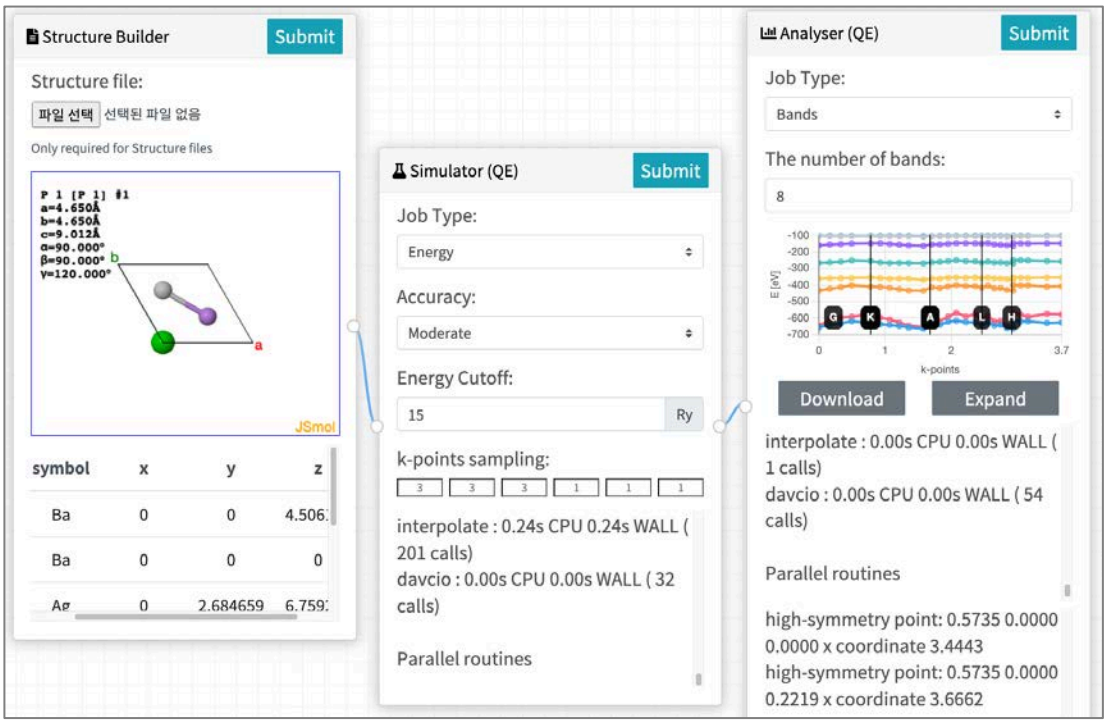

Fig. 9. DFT simulation using the user-defined data type

Based on the research, characteristic data types can be created for various heterogeneous data, and features based on data types can be automatically extracted for data-driven simulation or analysis.

6. EDA Service and Data Enrichment using AI Model

Table 4. EDA Service functions

Pipeline	Function	Description
Load	From Local	Loads local files from the user's PC
	From SDR	Loads data files from the Science Data Repository
	From Remote	Loads data files using an external FTP or API
Explore	All Data	Displays all read data
	Top-N	Displays N data from the top of the file
	Bottom-N	Displays N data from the bottom of the file
	Range	Shows data in the range (M, N) position
Clean	Missing data	Handles missing data
	Outliers	Handles outliers
Transform	Label Encoding	Transforms the category variables using a label encoding
	One-hot Encoding	Transforms the category variables using a one-hot encoding
	Standard Scaling	Transforms data using a standard scaling
	MinMax Scaling	Transforms data using a minmax scaling
Enrich	Expert-based	Performs human-based data enrichment
	Rule-based	Performs the built-in algorithmic approach
	AI-based	Loads and run the AI model

Select	Sampling	Extracts sample data
	Merging	Merges two files
	Split	Splits the data by a given ratio
Visualize	Heatmap	Visualizes data heatmap
	Missing bar	Visualizes missing data
	Box plot	Visualizes box plot
	Feature Importance	Calculates the feature importance and represents it as a graph
Save	To SDR	Saves a result file to the Science Data Repository

When using machine learning algorithms to analyze data, the most time-consuming and challenging part is cleaning and organizing data [42]. Even if we have the best AI training model, it will not perform well with poorly preprocessed data, and the accuracy of the model will decrease. The EDISON 2.0 data framework provides EDA services to support data preprocessing functions. Table 4 lists the functions supported by the EDA service [43,44]. Data preprocessing does not end with one or two functions. It is a long and tedious task that is performed by repeating a series of operations such as data loading, exploring, cleaning, transforming, enriching, selecting, and saving. During the data-processing pipeline, visualization checks are essential to ensure that the data have been appropriately cleaned. The EDA service simplifies this process by defining pipeline-like functions, including handling missing data and outliers, using techniques such as filling in constant values or removing rows or columns [45].

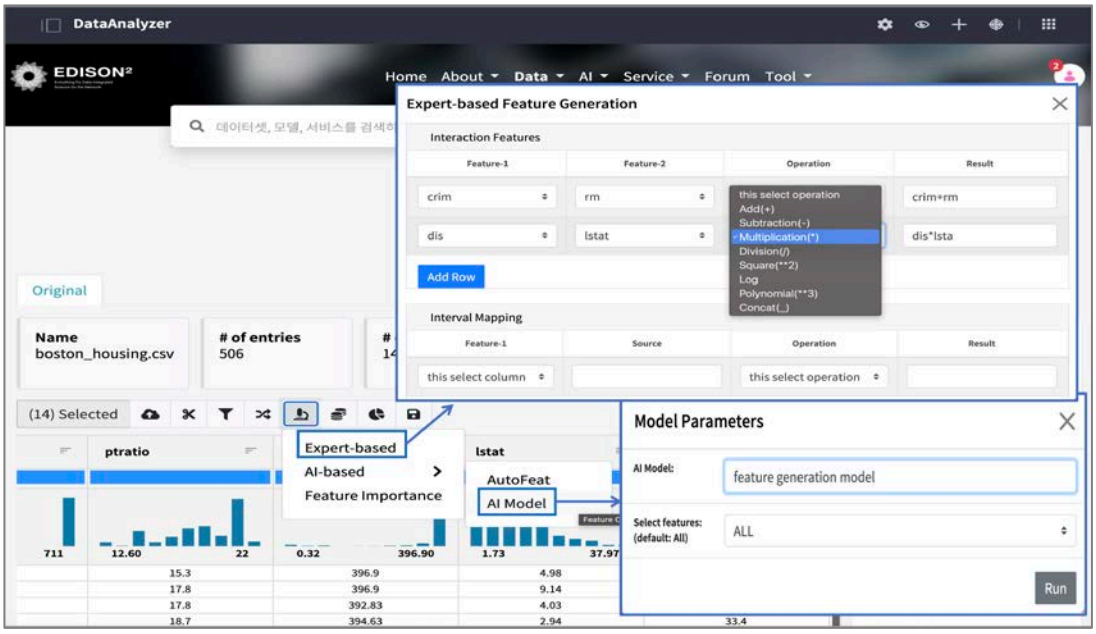

Fig. 10. Data-enrichment example

Data-driven interdisciplinary convergence research extracts and predicts insights from data by combining expertise in various fields. In particular, the unique meaning of individual data must be determined; however, as AI technology advances, accumulating a large amount of data and understanding the meaning of the entire data is more important. To this end, the EDISON 2.0 data framework provides an environment that can support data enrichment by applying the latest AI technology to large-scale data accumulated in a data lifecycle-based data

repository. Researchers can perform data enrichment during data preparation before full-scale software execution. Data enrichment can significantly improve the accuracy and value of data and yield more reliable research results. **Fig. 10** shows an example of performing feature generation among the data enrichment functions in the EDA service. This service provides three feature generation functions as follows:

***Expert-based function*:** This is a human-based feature generation method. The person who knows the data best, that is, the expert, selects the features that need to be enriched in the data and the parameter options, and then generates new features.

***Rule-based function*:** It is a built-in implementation of a well-known machine learning algorithm. This feature can be used when we do not know the data well or want to use a machine learning algorithm to gain new insights. The current implementation uses the AutoFeat package [46].

***AI model-based function*:** This generates features using AI models registered in the AI framework of EDISON 2.0. When researchers register high-performing AI models in an AI framework, users of the EDA service can load those models and directly apply them to their own data. This has a significant advantage of allowing easy and rapid utilization of the latest technologies.

New features created through these three methods can be evaluated for their suitability in the model using the feature importance function. Using the feature importance function, we can determine which features are important in the newly created data and obtain the accuracy or root mean squared error (RMSE) for the machine learning model. The machine learning algorithm implemented in feature importance uses LightGBM [47].

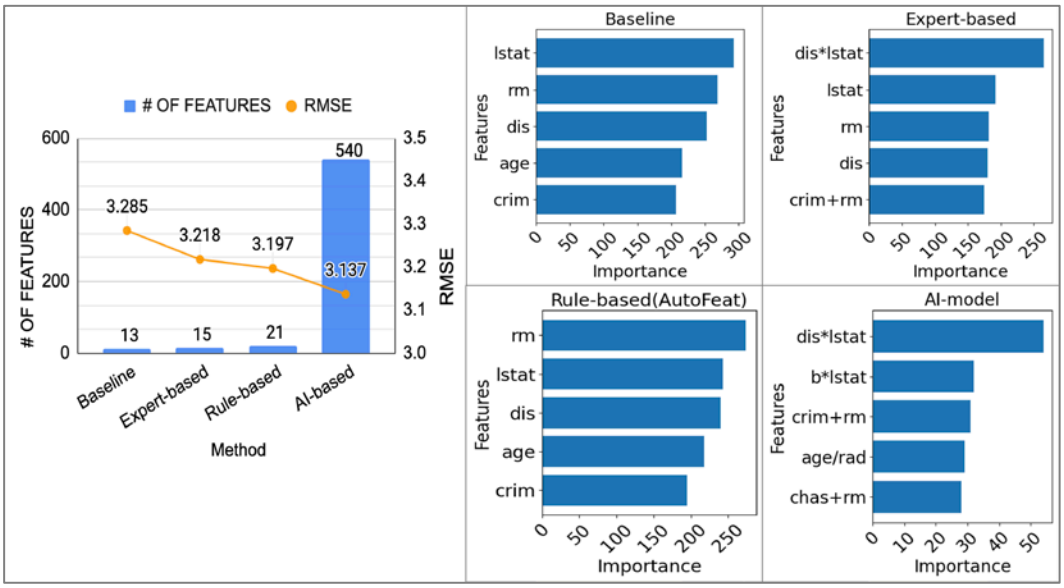

Fig. 11. Data enrichment example

Fig. 11 show the results of using the above three data enrichment techniques using the Boston housing dataset [48,49]. The baseline method shows the state when no data enrichment is applied. In this case, the number of features in the data is 13. The expert-based function has 15 features, rule-based function (AutoFeat algorithm) has 21 features, and AI model-based function has 540 features, which is more than 40 times more features than that of the baseline. Because we are dealing with a regression problem to predict house prices, we use the RMSE

score [50]. As can be observed in the figure, the RMSE score decreases as the number of features increases. This indicates that the AI model-based method has better data. However, results may vary depending on the situation. The right-side image demonstrates the feature importance of each method, aiding users in discerning vital data features for problem-solving. Consequently, the EDISON 2.0 EDA service facilitates data enrichment using AI models, thereby enabling the creation of superior datasets for AI-model training.

Fig. 12 illustrates the interaction of the data framework with the AI and simulation frameworks. The data framework is enhanced by the ability to import AI models from the AI framework's registry, expanding its capacity for AI-based data enrichment functions such as auto-extraction, derived data generation, auto-classification, data labeling, and more. Additionally, the data framework can operate not just on a local server but also on the EDISON cloud, which is currently being developed. The EDISON scheduler facilitates this, enabling processing of large data volumes using EDISON cloud clusters.

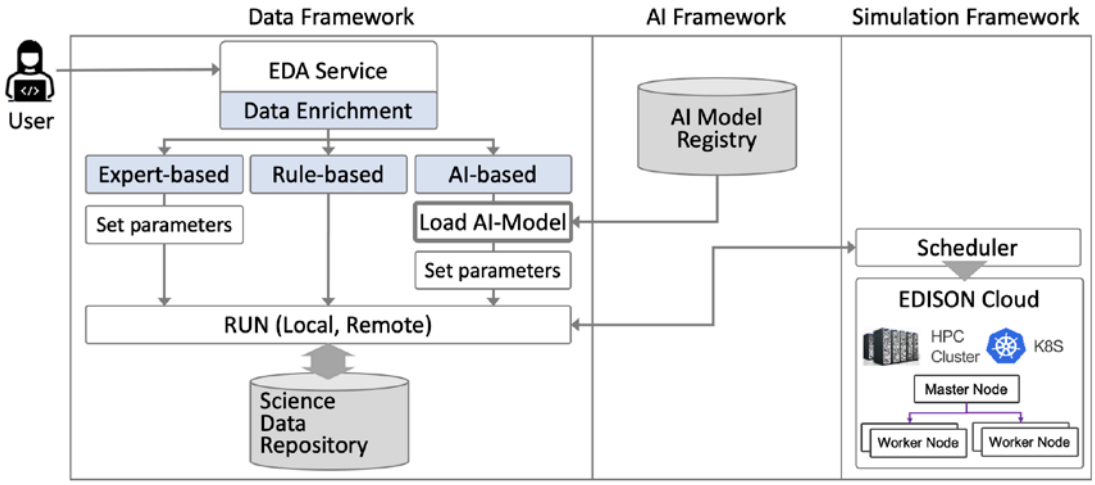

Fig. 12. Interactions between frameworks example

7. Conclusion

As the importance of data increases and AI technology advances, the research paradigm is changing and various digital platforms are being created to meet these needs. In the field of computational science with a large amount of computer computation, various platforms are being used to develop models and simulations through large-scale parallel computation. The platform makes it easy for researchers to use large computing resources and allows them to share and collaborate on the large-scale data, techniques, and tools used in their research. In particular, the initial simulation and simulation-driven platforms are changing and developing into data-driven and AI-driven platforms.

HUBzero was created as a web-based science and engineering collaboration platform and developed the PURR platform for data sharing and management and the MyGeoHub platform for enhancing metadata and file management capabilities. The Materials Project was created to distribute computational results and analysis tools for materials data, and the MPContribs platform was developed to improve the data integration and search capabilities. AiiDA was created as a platform for research reproducibility by tracking scientific workflows, and the Materials Cloud was developed to share, search, and visualize everything tracked by AiiDA.

EDISON 1.0 was developed as an online simulation platform created for educational and research purposes in scientific and engineering professional applications to form and enable a large computational science community. Recently, however, there has been a growing demand for a shift from a simulation-centered education platform to a data- and AI-centered research platform. In this study, the most central data framework was proposed to develop EDISON 1.0, which is simulation-centered, into EDISON 2.0, which is a data-centered research platform. First, the proposed framework provides a data repository suitable for the data lifecycle. Researchers have the advantage of increasing the reproducibility of their research by managing the data generated at each stage. Second, it provides a heterogeneous data-integration environment. A new data model for the integration, management, search, and utilization of heterogeneous data and an interdisciplinary convergence research environment using data-centered heterogeneous data were proposed. Finally, it supports the EDA service and data enrichment using AI model. It utilizes the latest AI technology to generate new derivative data from accumulated data and provides an environment where data can be easily explored and refined through EDA services. This provides researchers without IT expertise the opportunity to quickly process data and gain insights from data in other fields to conduct new and challenging research. The data framework of EDISON 2.0 will continue to develop its functions through data-driven interdisciplinary convergence research.

EDISON 2.0 excels at integrating and processing heterogeneous data, but data quality is limited by the data providers. Therefore, future work is to continuously improve data quality without relying on data providers. The purpose is to increase the reliability of research results by managing everything from data collection to predefined quality rules with an automated system.

References

[1] Wikipedia contributors, “Computational Science,” *Wikipedia*. [Online]. Available: https://en.wikipedia.org/wiki/Computational_science, Accessed on: Feb. 18, 2023.

[2] A. B. Shiflet and G. W. Shiflet, *Introduction to computational science: modeling and simulation for the sciences*, Princeton, NJ: University Press, 2014.

[3] L. Cao, “Data science: a comprehensive overview,” *ACM Computing Surveys (CSUR)*, vol. 50, no. 3, pp. 1-42, 2017. Article (CrossRef Link)

[4] C. O. Klingenberg, M. A. V. Borges, and J. A. V. Antunes Jr, “Industry 4.0 as a data-driven paradigm: a systematic literature review on technologies,” *Journal of Manufacturing Technology Management*, vol. 32, no. 3, pp. 570-592, 2021. Article (CrossRef Link)

[5] G. D. Bisol et al., “Perspectives on Open Science and scientific data sharing: an interdisciplinary workshop,” *Journal of Anthropological Sciences*, vol. 92, pp. 179-200, 2014. Article (CrossRef Link)

[6] C. Allen and D. M. A. Mehler, “Open science challenges, benefits and tips in early career and beyond,” *PLoS Biology*, vol. 17, no. 12, e3000246, 2019. Article (CrossRef Link)

[7] J. Wing, “Computational thinking’s influence on research and education for all,” *Italian Journal of Educational Technology*, vol. 25, no. 2, pp. 7-14, 2017. Article (CrossRef Link)

[8] Y. Zhao, I. Raicu, and I. Foster, "Scientific workflow systems for 21st century, new bottle or new wine?," *2008 IEEE Congress on Services-Part I*, 2008. Article (CrossRef Link)

[9] T. Mucha and T. Seppala, “Artificial Intelligence Platforms-A New Research Agenda for Digital Platform Economy,” *Social Science Electronic Publishing*, Feb. 2020. Article (CrossRef Link).

[10] A. Lavin et al., “Simulation intelligence: Towards a new generation of scientific methods,” *arXiv:2112.03235*, 2022. Article (CrossRef Link).

[11] D.-S. Jin, Y.-J. Jung, and Hoe-Kyung Jung, “EDISON platform to supporting education and integration research in computational science,” *Journal of the Korea Institute of Information and Communication Engineering*, vol. 16, no. 1, pp. 176-182, 2012. Article (CrossRef Link)
[12] J. Ma, J. R. Lee, K. Cho, and M. Park, “Design and implementation of information management tools for the EDISON open platform,” *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 11, no. 2, pp. 1089–1104, 2017. Article (CrossRef Link)
[13] S. Menken et al., *An introduction to interdisciplinary research: Theory and practice*, Amsterdam, The Netherlands: Amsterdam University Press, 2016.
[14] D. Sui and J. Coleman, “Convergence Research in the Age of Big Data: Team Science, Institutional Strategies, and Beyond,” *Merrill Series on The Research Mission of Public Universities*, pp. 23-35, 2019. Article (CrossRef Link)
[15] S. N. Goodman, D. Fanelli, and J. P. Ioannidis, “What does research reproducibility mean?,” *Science Translational Medicine*, vol. 8, no. 341, 2016. Article (CrossRef Link)
[16] O. E. Gundersen and S. Kjensmo, “State of the art: Reproducibility in artificial intelligence,” in *Proc. of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, pp. 1644-1651, 2018. Article (CrossRef Link)
[17] X. Chen and M. Jagerhorn, “Implementing FAIR Workflows along the research lifecycle,” *Procedia Computer Science*, vol. 211, pp. 83-92, 2022. Article (CrossRef Link)
[18] R. Kusters et al., "Interdisciplinary research in artificial intelligence: challenges and opportunities," *Frontiers in big data*, vol. 3, 2020. Article (CrossRef Link)
[19] A. Asadullah, I. Faik, and A. Kankanhalli, “Digital Platforms: A Review and Future Directions,” in *Proc. of the Pacific Asia Conference on Information Systems (PACIS)*, 248, 2018. Article (CrossRef Link)
[20] C. Bonina et al., “Digital platforms for development: Foundations and research agenda,” *Information Systems Journal*, vol. 31, no. 6, pp. 869-902, 2021. Article (CrossRef Link)
[21] M. De Reuver, C. Sørensen, and R. C. Basole, “The digital platform: a research agenda,” *Journal of information technology*, vol. 33, no. 2, pp. 124-135, 2018. Article (CrossRef Link)
[22] P. Das et al., “Amazon SageMaker Autopilot: a white box AutoML solution at scale,” in *Proc. of the 4th International Workshop on Data Management for End-to-End Machine Learning*, 2020. Article (CrossRef Link)
[23] M. Salvaris, D. Dean, and W. H. Tok, “Microsoft AI platform,” in *Deep Learning with Azure: Building and Deploying Artificial Intelligence Solutions on the Microsoft AI Platform*, 2018, pp. 79-98. Article (CrossRef Link)
[24] E. Bisong, “An overview of google cloud platform services,” in *Building Machine Learning and Deep Learning Models on Google Cloud Platform: A Comprehensive Guide for Beginners*, 2019, pp. 7-10. Article (CrossRef Link)
[25] M. McLennan and R. Kennell, “HUBzero: a platform for dissemination and collaboration in computational science and engineering,” *Computing in Science and Engineering*, vol. 12, no. 2, pp. 48-53, 2010. Article (CrossRef Link)
[26] S. Gesing, M. Zentner, S. Clark, C. Stirm, and B. Haley, “HUBzero®: Novel Concepts Applied to Established Computing Infrastructures to Address Communities' Needs,” in *Proc. of the Practice and Experience in Advanced Research Computing on Rise of the Machines (learning)*, pp. 1-7, 2019. Article (CrossRef Link)
[27] C. Dearborn, A. J. Barton, and N. Harmeyer, “The Purdue University Research Repository,” *OCLC Systems & Services*, vol. 30, no. 1, pp. 15-27, 2014. Article (CrossRef Link)
[28] R. Kalyanam et al., “MyGeoHub-A sustainable and evolving geospatial science gateway,” *Future Generation Computer Systems*, vol. 94, pp. 820–832, May 2019. Article (CrossRef Link)
[29] R. Kalyanam, L. Zhao, R. Campbell, and C. Song, "Data-driven Collaboration Environments: Integrating HUBZero and iRODS," 2016. Article (CrossRef Link)
[30] B. Blaiszik et al., “The Materials Data Facility: Data Services to Advance Materials Science Research,” *Journal of the Minerals, Metals and Materials Society*, vol. 68, no. 8, pp. 2045–2052, Jul. 2016. Article (CrossRef Link)

[31] A. Jain et al., "The Materials Project: Accelerating Materials Design Through Theory-Driven Data and Tools," in *Springer International Publishing eBooks*, Jan. 2020, pp. 1751–1784. Article (CrossRef Link)
[32] P. Huck et al., "A Community Contribution Framework for Sharing Materials Data with Materials Project," in *Proc. of 2015 IEEE 11th International Conference on e-Science*, pp. 535-541, 2015. Article (CrossRef Link)
[33] G. Pizzi et al., "AiiDA: automated interactive infrastructure and database for computational science," *Computational Materials Science*, vol. 111, pp. 218–230, 2016. Article (CrossRef Link)
[34] J. R. Wilcox et al., "Verdi: a framework for implementing and formally verifying distributed systems," in *Proc. of the 36th ACM SIGPLAN Conference on Programming Language Design and Implementation*, 2015. Article (CrossRef Link)
[35] L. Talirz et al., "Materials Cloud, a platform for open computational science," *Scientific Data*, vol. 7, no. 1, 2020. Article (CrossRef Link)
[36] S. Ahn, J. Lee, J.-S. Kim, and J. R. Lee, "EDISON-DATA: A flexible and extensible platform for processing and analysis of computational science data," *Software - Practice and Experience*, vol. 49, no. 10, pp. 1509-1530, 2019. Article (CrossRef Link)
[37] "Research lifecycle," *Research Support at Harvard*. [Online]. Available: https://researchsupport.harvard.edu/research-lifecycle, Accessed on: Feb. 18, 2023.
[38] J. L. Faundeen et al., "The United States Geological Survey Science Data Lifecycle Model," *Open-file Report*, 2014. Article (CrossRef Link)
[39] M. E. Arass and N. Souissi, "Data Lifecycle: From Big Data to SmartData," in *Proc. of 2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, 2018. Article (CrossRef Link)
[40] S. Weibel and T. Koch, "The Dublin Core Metadata Initiative," *D-lib Magazine*, vol. 6, no. 12, 2000. Article (CrossRef Link)
[41] S. Kirklin et al., "The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies," *NPJ Computational Materials*, vol. 1, no. 1, Dec. 2015. Article (CrossRef Link)
[42] Q. Yao et al., "Taking human out of learning applications: A survey on automated machine learning," *arXiv:1810.13306*, 2019. Article (CrossRef Link)
[43] K. Sahoo, A. K. Samal, J. Pramanik, and S. K. Pani, "Exploratory data analysis using Python," *International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 12, pp. 4727-4735, 2019. Article (CrossRef Link)
[44] X. Chu et al., "Data cleaning: Overview and emerging challenges," in *Proc. of the 2016 international conference on management of data*, pp. 2201-2206, 2016. Article (CrossRef Link)
[45] "Interquartile range," [Online]. Available: https://en.wikipedia.org/wiki/Interquartile_range, Accessed on: Feb. 18, 2023.
[46] F. Horn, R. Pack, and M. Rieger, "The autofeat python library for automated feature engineering and selection," in *Machine Learning and Knowledge Discovery in Databases: International Workshops of ECML PKDD 2019*, Würzburg, Germany, September 16–20, 2019, Proceedings, Part I. Springer International Publishing, 2020. Article (CrossRef Link)
[47] G. Ke et al., "Lightgbm: A highly efficient gradient boosting decision tree," in *Proc. of the 31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017. Article (CrossRef Link)
[48] D. Harrison and D. Rubinfeld, "UCI Machine Learning Repository,". [Online]. Available: http://lib.stat.cmu.edu/datasets/boston, Accessed on: Feb. 18, 2023.
[49] S. Sanyal et al., "Boston house price prediction using regression models," in *Proc. of 2022 2nd International Conference on Intelligent Technologies (CONIT)*, IEEE, 2022. Article (CrossRef Link)
[50] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, 2021. Article (CrossRef Link)

Sunggeun Han received the M.S. degree in computer engineering from Jeonbuk National University in 2000. He is currently working as a Principal Researcher at Korea Institute of Science and Technology Information (KISTI) in South Korea. His research interests include digital platform development, data science, and artificial intelligence in HPC convergence area.

Jaegwang Lee received the M.S. degree and Ph.D. in Computer Engineering from Hannam University in 2014 and 2018, respectively. Since October 2018, he has been working in the HPC Convergence R&D Platform Department at Korea Institute of Science and Technology Information (KISTI). His research interests are computer networks, artificial intelligence and R&D platforms.

Inho Jeon received his B.S., Ph.D. degree in electric engineering from Kwangwoon University. He is a senior researcher at National Institute of Supercomputing and Networking, Korea Institute of Science and Technology Information (KISTI). His research interests include cloud service, science gateway and open science.

Jeongcheol Lee received the B.S., M.S. and Ph.D. degrees in Computer Engineering from Chungnam National University, Daejeon, Korea, in 2008, 2010 and 2014, respectively. He is currently working as a principal researcher in Korea Institute of Science Technology and Information (KISTI) in South Korea. He is interested in AI, data framework, and computational science in HPC convergence area.

Hoon Choi received a BS in computer engineering from the Seoul National University, Republic of Korea in 1983, and his MS and PhD in computer science from Duke University in 1990 and 1993, respectively. From 1983 to 1996, He was a senior member of technical staff at ETRI, Korea where he worked on LAN, broadband ISDN and high-speed network systems. He is with Chungnam National University since 1996 and he worked for the Advanced Network Technologies Division of NIST, U.S.A. as a guest researcher in 2000. His recent research interests focuses on the middleware for mobile distributed computing and autonomic computing.