A Design and Implementation of a Platform for Generating On-Demand Training Datasets

Jooyoung Lee
Intelligent Network Research Section
Electronics and Telecommunications Research Institue
Daejeon, South Korea
joolee@etri.re.kr

Chunglae Cho
Intelligent Network Research Section
Electronics and Telecommunications Research Institue
Daejeon, South Korea
clcho@etri.re.kr

Hongseok Jeon
Intelligent Network Research Section
Electronics and Telecommunications Research Institue
Daejeon, South Korea
jeonhs@etri.re.kr

Seunghyun Yoon
Intelligent Network Research Section
Electronics and Telecommunications Research Institue
Daejeon, South Korea
shpyoon@etri.re.kr

***Abstract*—This paper introduces a platform designed to process extensive network data gathered from various sources according to user-defined order specifications, resulting in the creation of training datasets. The proposed platform enables users to obtain their learning datasets through a service, eliminating the need to construct their own preprocessing systems when generating training datasets from raw data for AI/ML learning. Particularly, when dealing with substantial data volumes, users can efficiently process data through distributed processing on the service's provided big data platform. Additionally, enabling on-demand data processing makes it possible to generate datasets suitable for the learning models users intend to develop, thereby enhancing productivity during the training process of artificial intelligence models.**

Keywords—On-demand dataset, Network dataset, AIML Data preprocessing platform, Bigdata processing, Insights

I. Introduction

With the increasing complexity of network infrastructures in recent years, the need for network management based on artificial intelligence technology has grown significantly. As a result, securing high-quality datasets for training AI/ML models has become an essential requirement. The process of acquiring such datasets involves refining and organizing collected raw data into a suitable format for learning. In this context, network data, which is unstructured and accumulates rapidly in substantial quantities, demands typical big data processing. Furthermore, to prevent overfitting of artificial intelligence models to specific datasets and ensure generalization, training datasets should reflect the traffic characteristics originating from various regions and the fluctuations in traffic patterns over time.

The task of generating training datasets to align with the characteristics of network data demands a significant investment of time and resources. Currently, it involves researchers individually constructing the necessary resources and developing processing methods tailored to the specific artificial intelligence models they are working on. However, as mentioned earlier, acquiring and managing the required support for processing extensive data is not only costly but also far from straightforward. To address these challenges, a platform based on a service-oriented approach that refines and processes raw data according to user orders to generate training datasets is necessary.

II. Related Works

To realize autonomous networks through intelligent functionalities and control, it's crucial to acquire diverse network datasets that comprehensively represent various characteristics specific to each application domain. These datasets will serve as the foundation for developing intelligent network functions and control technologies that leverage AI/ML, ultimately leading to enhanced predictability, autonomy, and optimization in communication and networking.

At present, the types of data primarily used for the development of AI/ML models in the field of network are categorized and presented in [1]. As most network infrastructures are reliant on communication service providers, cloud operators, application service providers, and the like, obtaining data suitable for AI/ML applications poses practical challenges. Consequently, many studies resort to utilizing publicly available datasets such as the UNSW-NB15 dataset [2], MAWI dataset [3], TOTEM dataset [4], CAIDA2007 [5], KDDCUP [6], KDDCUP99 [7], NSL-KDD [7], CICDDoS2019 [8], CICIDS2017 [9], and ISCX datasets (ISCXVPN2016, ISCXTor2016) [10]. However, these datasets might not comprehensively capture the characteristics of contemporary network traffic due to their age.

An example of a service that collects and processes network data to create datasets is Huawei's NAIE (Network AI Engine) AI Services [11]. This service offers dataset browsing, querying, and subscription services for communication network and device datasets. The dataset types includes wireless access training dataset, fixed access training dataset, bearer network training dataset, core network training dataset, and more.

III. Design of Platform for Generating Order-Based Training Dataset

The paper aims to address the limitations of conventional technologies by introducing a platform designed to generate artificial intelligence training data, which is utilized throughout the entire lifecycle of artificial intelligence models.

This work was supported by the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. 2021-0-00851, On-demand data based network intelligence framework technology development).

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

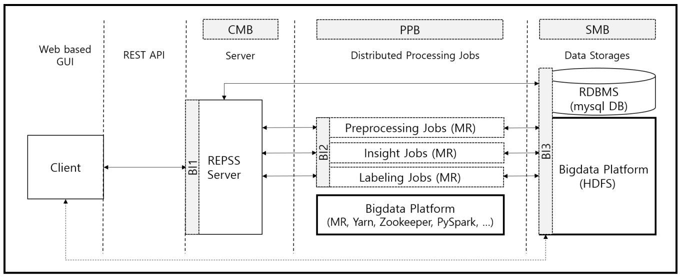


Fig. 1. Architecture of the proposed platform

The platform offers an on-demand data preprocessing service designed to generate network AI/ML datasets by processing diverse raw data according to specific objectives. It receives preprocessing requests from users for large-scale collected data and conducts preprocessing based on the specified requirements using a distributed data processing big data platform. This leads to the production of refined Net-AI data, which is subsequently provided to users through an order-based data preprocessing service.

A. Architecture

The proposed platform offers various services, including executing preprocessing orders, managing such orders, and serving as a repository for both raw data and refined AI/ML data. To achieve this, it consists of three main components: Control Block (CMB), Preprocessing Block (PPB), and Repository Block (SMB). Each block operates as shown in Fig. 1 and is responsible for the following functions:

- CMB is built as a web server, receiving data preprocessing and data repository query requests from user clients and providing processing results. When multiple preprocessing steps are defined in an order, it performs pipelining for data processing. Requests and responses for each service are conducted through an HTTP-based OpenAPI (BI1).

- PPB consists of function-specific library modules for preprocessing, insights, and labeling, as well as distributed processing process execution modules. Users' preprocessing code, described using the interface (BI2) of library modules, is executed as a distributed process on the big data platform.

- SBM is composed of a distributed file system for large-scale file storage and a relational database system for structured data management. Operations such as reading, writing, and deletion for each storage system are performed through the interfaces (BI3) provided by each storage system.

Additionally, a web-based GUI is provided to offer visualized execution results. In Fig. 1, the bold lines within the diagram represent the utilization of a big data platform, which is built upon an open-source framework that supports distributed processing.

B. Open APIs

In this paper, BI1 interfaces of the platform are defined as REST APIs following the standards of the OpenAPI Initiative, allowing remote systems to make API calls using HTTP or HTTP-like protocols.

BI1 interface provides the following four types of service APIs, which can be utilized based on their purpose:

- Data Preprocessing APIs: These interfaces are used to preprocess raw data based on requests specified in the order specifications, with the goal of generating AI/ML datasets.

- Repository Data Management APIs: These interfaces are employed to query or delete datasets and their associated metadata stored in the repository.

- Dataset Insight APIs: These interfaces execute data processing based on order specifications, enabling data visualization to provide insights into AI/ML datasets.

- Dataset Sharing APIs: These interfaces are utilized for downloading AI/ML datasets and retrieving metadata associated with them.

C. OrderForm

The platform offers a data schema named "OrderForm," enabling users to define their specific requirements while requesting dataset creation. Within the OrderForm, as depicted in Table 1, users outline the source data and the intended output data for generating the training dataset. This schema also encompasses programing codes for data processing. Multiple processing methods can be indicated for data manipulation, resulting in the creation of a processing pipeline that systematically applies these methods to the input data.

Currently, the data processing methods can be specified using Python and PySpark. The resulting output datasets can be saved in formats such as CSV, AVRO, and JSON. Additionally, further processing of the generated datasets can be conducted to extract insights, enabling the visualization of graph charts based on the processed outcomes.

TABLE I. ORDERFORM SPECIFICATIONS USED IN THE PROPOSED PLATFORM

Elements	Data Type (in yaml)	Descriptions
orderName	string	Name of order
description	string	Description of order
inputFormat	string	Method to retrieve input data
input	oneOf : string($uri), array of URL, array of DataID	Path of input data
output	string($uri)	Path of output data
prepList*	PreSpecification	List of preprocessing specs.
dataSchemaDesc	DataSchemaDesc	Fields and descriptions of preprocessed data schema
outputFileFormat	OutputFileFormat	File format of output dataset
saveTempResult	boolean	Storage of intermediate preprocessing data
callbackURI	string($uri)	Notification address for preprocessing completion

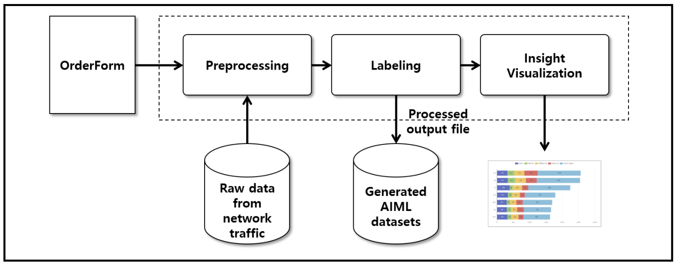


Fig. 2. A work flow to generate a dataset as specified in an OrderForm

IV. CASE STUDY

In this section, a scenario is provided to configure simple preprocessing and insight requirements for raw data.

A. Multiple Orders

- Preprocessing: The preprocessing requirement involves replacing empty fields in the input data with an integer value of '0'. This preprocessing is accomplished by providing Mapper code written in Python to perform distributed processing through Hadoop's MapReduce.

- Labeling: The labeling requirement is to generate labels for "Classification" and "Prediction" from the preprocessed data. The example order form includes PySpark code for this labeling task, where PySpark internally executes distributed processing based on Hadoop. The labeled data is stored in the repository as the final output.

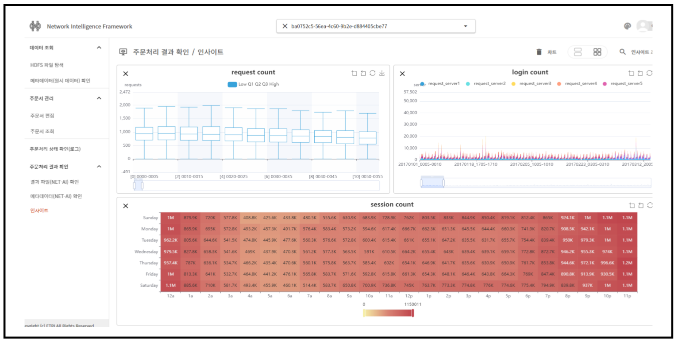
- Insight: For this sample's insight requirement, the goal is to reorganize two years' worth of session request count data stored at 5-minute intervals into day-of-the-week and hour-of-the-day categories. This allows for an intuitive visualization of when session requests were most frequent on which days and at which times over the two-year period. The example order form includes Hadoop Mapper and Reducer code written in Python for processing this insight-related data.

B. Preprocessed Results

Fig. 3 displays the visualization of statistical data for datasets generated from processing the order requirements specified in the previous scenario. The visualization includes stacked bar charts, box plots, and heatmaps. Users can utilize these visualizations to examine the distribution and characteristics of the training dataset. Additionally, they can make judgments related to labeling values necessary for training and threshold values for Classification. Visualizing large-scale data is crucial for gaining insights. While Huawei's Data Lake Insight [12], as one of the network dataset provisioning services, offers large-scale data processing and analysis capabilities, it differs from the platform proposed in this paper in that it does not provide methods for data visualization.

V. CONCLUSION

In this paper, we introduced a platform for generating AI/ML training datasets according to user orders using diverse network data collected from various sources. This order-based training dataset generation platform offers the advantage of enabling users to create suitable datasets for their learning models without the need to directly build preprocessing systems. This is achieved through a service-oriented approach. Especially when dealing with large amounts of data, users can efficiently process data through distributed processing on the provided big data platform. The proposed platform is expected to enhance productivity and efficiency during the training process of AI models. Going forward, we have plans to release the network datasets generated through this platform to the public.



	12a	1a	2a	3a	4a	5a	6a	7a	8a	9a	10a	11a	12p	1p	2p	3p	4p	5p	6p	7p	8p	9p	10p	11p
Sunday	1M	879.9K	720K	577.8K	408.8K	425.6K	433.8K	480.5K	555.6K	630.9K	683.9K	728.9K	762K	803.5K	833K	844.9K	850.4K	819.1K	812.4K	865K	924.1K	1M	1.1M	1.1M
Monday	1M	865.9K	695K	572.8K	493.2K	457.3K	491.7K	576.4K	583.4K	573.2K	594.6K	617.4K	666.7K	662.3K	651.3K	645.5K	644.4K	660.3K	741.9K	820.7K	908.5K	942.1K	1M	1.1M
Tuesday	962.2K	805.6K	644.6K	541.5K	474.8K	445.9K	477.6K	560.3K	576.6K	572.8K	600.4K	615.4K	661K	655.1K	647.2K	635.5K	631.7K	655.7K	754.4K	839.4K	950K	979.3K	1M	1.1M
Wednesday	979.5K	827.8K	658.3K	541.6K	469K	437.9K	470.3K	561.2K	577.3K	563.5K	591K	610.5K	664.2K	655.4K	643K	639.4K	639.1K	659.1K	772.8K	872.7K	946.2K	955.3K	974K	1.1M
Thursday	957.4K	787K	636.1K	534.7K	466.2K	435.4K	470.6K	560.1K	575.8K	563.7K	585.4K	602K	654.1K	646.9K	641.7K	635.6K	630.9K	650.9K	761.7K	853.8K	944.6K	972.1K	996.6K	1.2M
Friday	1M	813.3K	641K	532.7K	464.8K	441.2K	476.1K	565.8K	583.7K	571.6K	592.8K	615.8K	661.3K	654.3K	648.1K	646.4K	643.8K	664.3K	769K	847.4K	890.8K	913.9K	930.5K	1.1M
Saturday	1.1M	885.6K	710K	581.7K	493.4K	455.9K	460.1K	514.4K	583.7K	650.8K	700.9K	736.8K	745K	763.7K	773.3K	774.8K	776K	774.6K	775.4K	794.9K	839.8K	937K	1M	1.1M

Fig. 3. Case study results: Visualizing statistical data of generated dataset

REFERENCES

[1] Jooyoung Lee, et al. "Survey on Artificial Intelligence & Machine Learning Models and Datasets for Network Intelligence." The Journal of Korean Institute of Communications and Information Sciences, VOL. 47, NO. 4, April 2022, pp 625-643. (In Korean)

[2] N. Moustafa, The UNSW-NB15 Dataset, Retrieved Nov. 28, 2021, from https://research.unsw.edu.au/projects/unsw-nb15-dataset

[3] Y. Zhong, et al., "HELAD: A novel network anomaly detection model based on heterogeneous ensemble learning," Comput. Netw., vol. 169, 107049, Mar. 2020.

[4] S. Balon, J. Lepropre, O. Delcourt, F. Skivee, G. Leduc, TOTEM Project: TOolbox for Traffic Engineering Methods, Retrieved Nov. 28, 2021, from http://totem.run.montefiore.ulg.ac.be/

[5] R. K. Deka, D. K. Bhattacharyya, J. K. Kalita, "Active learning to detect DDoS attack using ranked features," Comput. Commun., vol. 145, pp. 203-222, Sep. 2019.

[6] J.-F. Cui, H. Xia, R. Zhang, B.-X. Hu, X.-G. Cheng, "Optimization scheme for intrusion detection scheme GBDT in edge computing center," Comput. Commun., vol. 168, pp. 136-145, Feb. 2021.

[7] S. Gamage, J. Samarabandu, "Deep learning methods in network intrusion detection: A survey and an objective comparison," J. Netw. Comput. Appl., vol. 169, 102767, Nov. 2020.

[8] J. P. A. Maranhão, J. P. C. L. da Costa, E. Javidi, C. A. B. de Andrade, R. T. de Sousa Jr., "Tensor based framework for distributed denial of service attack detection," J. Netw. Comput. Appl., vol. 174, 102894, Jan. 2021.

[9] S. D. Çakmakçı, T. Kemmerich, T. Ahmed, N. Baykal, "Online DDoS attack detection using Mahalanobis distance and kernel-based learning algorithm," J. Netw. Comput. Appl., vol. 168, 102756, Oct. 2020.

[10] Lin, Xinjie, et al. "Et-bert: A contextualized datagram representation with pre-training transformers for encrypted traffic classification.", Proceedings of the ACM Web Conference 2022

[11] Huawei Technologies co. LTD., "Dataset Service V200R021C50", from https://support.huaweicloud.com/en-us/usermanual-naie-dataset/Dataset%20Service.pdf

[12] Huawei Technologies co. LTD., "Product Introduction: Data Lake Insight", from https://support.huaweicloud.com/en-us/productdesc-dli/dli-productdesc.pdf