# Designing Data Pipeline for Network Data Management in Digital Twin Network Environment

Hyeju Shin
*Department of ICT Convergence System Engineering*
*Chonnam National University*
Gwangju, Korea
sinhye102@gmail.com

Seungmin Oh
*Department of ICT Convergence System Engineering*
*Chonnam National University*
Gwangju, Korea
osm5252kr@gmail.com

Jihoon Lee
*Department of ICT Convergence System Engineering*
*Chonnam National University*
Gwangju, Korea
ghost6268@gmail.com

Gwangmoo Chung
*Department of ICT Convergence System Engineering*
*Chonnam National University*
Gwangju, Korea
chungkm1250@gmail.com

Jinsul Kim*
*Department of ICT Convergence System Engineering*
*Chonnam National University*
Gwangju, Korea
jsworld@jnu.ac.kr

***Abstract*—With the advancement of Internet of Things (IoT) and 5G technology, new network services such as Virtual Reality/Augmented Reality (VR/AR) are emerging, leading to the expansion of network scale and increasing network traffic. As a solution to the increasing complexity of network operation and maintenance, Digital Twin Network (DTN) technology has been introduced. In order to provide visibility into various aspects for network management, observation and analysis of network data are necessary. Therefore, this paper presents the types of data available in the digital twin network and the lifecycle process of data pipelines. This work will help to effectively manage data internally when implementing the Digital Twin Network in the future.**

***Keywords—digital twin network, network management, data pipeline***

## I. INTRODUCTION

With the development of Internet of Things (IoT), 5G, and cloud computing technologies, new network services such as Virtual Reality/Augmented Reality (VR/AR) and metaverse have emerged, and the network scale is increasingly expanding and the network load is increasing. In particular, due to the high reliability requirements of network operations, high network failure costs, and expensive testing costs, network changes often affect the entire network, making it increasingly difficult to deploy new technologies [1]. To address this, Digital Twin Network (DTN) or Network Digital Twin (NDT) technology has emerged [2]. It combines Digital Twin (DT) technology with communication networks to support network planning, construction, maintenance, and optimization, and is expected to be an important resource for network management and eventually autonomous networks.

In terms of network management, in the digital era, with the increasing complexity of computer networks, traditional device-centric network monitoring is not scalable and cannot provide sufficient visibility for applications [3], which requires observation and analysis of network data to correlate network availability and performance and link them to the state of applications. Therefore, this paper proposes an architecture of a data-centric digital twin network for flexible and scalable management of communication networks.

The rest of the paper is organized as follows. In Section 2, we survey and describe related work. In Section 3, we present the classification of available data and propose a network data pipeline structure using DTN for network management. Finally, Section 4 presents the conclusions along with future work.

## II. RELATED WORKS

### A. Internet Research Task Force (IRTF)

The Internet Research Task Force (IRTF) is a group that focuses on long-term research issues related to the Internet, working on topics related to Internet protocols, applications, architecture, and technology. The IRTF's Network Management Research Group (NMRG) is currently prioritizing research on three topics, including self-driving/-managing networks and artificial intelligence in network management [4]. As network operations and maintenance become more complex, the NMRG is discussing in an Internet Draft the application of digital twin technology to networks for comprehensive data-driven network infrastructure management throughout the entire network lifecycle [5]. The draft discusses the concept of a DTN and a reference architecture. A DTN should include four key elements: data, mappings, models, and interfaces, which are used to analyze, diagnose, emulate, and control real-world networks. Based on the definition of the technical elements, the resulting DTN architecture is divided into three layers: an application layer, a digital twin layer, and a physical network layer, as shown in Figure 1.

### B. Digital Twin Network (DTN)

DTN is a DT for communication network platforms that has recently emerged and is being standardized [5-6]. DTN can be

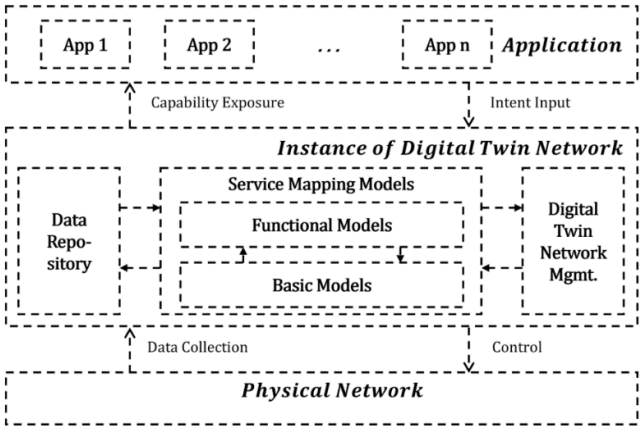


Fig. 1. IRTF's referemce architecture of DTN.

built by applying DT technology to map actual network equipment to virtual twin entities. The DTN architecture consists of three layers: the physical network layer, the digital twin network layer, and the network application layer. The physical network layer exchanges data information and network control information with the DTN layer. The DTN layer includes three main subsystems: data repository, service mapping model, and digital twin network management. The network application layer communicates the user's intent to the DTN layer to configure the DTN to perform appropriate tasks for the scenario.

DTN can be used to develop various network applications and evaluate specific policies before deploying updated configurations to the actual network. It can also be used to tune the network to optimize performance and easily analyze scenarios that are difficult to test in the real network, similar to real services [5]. By using DTN as an extension platform for network simulation, intelligent and efficient network management can be achieved.

## III. Network Data Management in DTN Environment

This section introduces the types of data available within the DTN for managing network data.

### A. System Architecture

The data pipeline is designed based on the DTN architecture shown in Figure 2 [7]. The data repository of the DTN layer includes data collector, data management, data storage, and data service. The service mapping model consists of a basic model and a functional model. These elements are configured to suit the user's intent through DTN management.

### B. Classification of Network Management Data

According to reference [3], the types of data that can be collected for Network Management (NM) can be classified into log/event data, metric data, flow data, packet data, network configuration information data, and forwarding/routing/path data. In addition, network data can be classified by type into network device data, network traffic data, and network performance data [8]. This information is summarized in Table 1, along with the available protocol candidates for carrying this data.

TABLE I. Available Data Types and Candidate Protocols

| *Data Type* | *Purpose* | *Data classification* | *Candidate Protocol* |
|---|---|---|---|
| Log/event data | Local equipment changes and event recognition. | Network device data | SNMP, Netconf |
| Metric data | Collecting performance and usage over time. | Network performance data | Telemetry |
| Flow data | Metadata for conversations between network endpoints. | Network traffic data | IPFIX, NetFlow |
| Packet data | Collecting entire conversation contents for diagnostic and security purposes. | Network traffic data | Netstream, NetFlow |
| Network configuratoin data | Capturing device configuration change management. | Network device data | Netconf |
| Forwarding/routing/path data | Optimizing traffic flow between network endpoints. | Network device data | SNMP, Netconf |

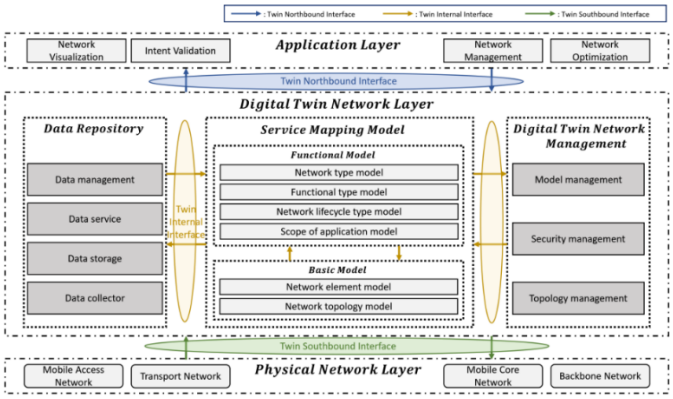


Fig. 2. Detailed overview of DTN architecture.

Network device data refers to data related to network elements, including geometric data, status data, event data, and topology data. Network performance data refers to data that includes performance data of various networks, such as transmission networks and data center networks, including network delay, packet loss, jitter, and bandwidth. Network traffic data refers to traffic data executed in the network, including packet length, traffic 5-tuple, and priority.

### C. Comparison of Data Used in Each Paper

Efficiently operating and managing networks requires the use of diverse network data. The accuracy of artificial intelligence models is significantly impacted by the quality and quantity of the data used. Therefore, it is crucial to understand the data's characteristics and preprocess, store, and transform it accordingly. To achieve this, we compared previous studies that utilized network data for network management, following the classification of network data types mentioned earlier. Table 2 provides a comprehensive comparison of input and output data, model types, and problem scope.

TABLE II. COMPARISON OF DATA AND MODELS USED IN EACH PAPER

| *Papers* | *Input/Output Data* | *Models* | *Problem Scope* |
|---|---|---|---|
| [9] | I: topology, edge&node features<br>O: classify the instruction | GraphSAGE | Instruction detection |
| [10] | I: topology types, botnet size<br>O: botnet detect | GCN | Botnet detection |
| [11] | I: topology, routing table, traffic matrix<br>O: delay, jitter, loss per src-dst | GNN | Src-dst KPI prediction |
| [12] | I: traffic patterns<br>O: traffic per horizon duration | GCN | Traffic forecasting in wide area networks |
| [13] | I: traffic, buffer size, topology, queue policy, routing scheme, etc.<br>O: path-level delay, flow-level delay, througnput | GNN | Network performance estimation |
| [14] | I: network status, flow table, QoS params(e.g., FCT, node through-put)<br>O: routing path | DQN, DDPG | Routing optimization |

## IV. DATA PIPELINE FOR NETWORK DATA MANAGEMENT IN DTN ENVIRONMENT

This section introduces a data pipeline within the DTN for managing network data.

### *A. Data Pipeline Architecture*

There were various structural proposals for effectively using DTN. However, there was no proposal for a data pipeline based on data structuring information along with the types of data that can be used within DTN. The configuration of a pipeline based on data types and structured information is important for effective network data management within the DTN. Fig. 3. illustrates the process of collecting and processing data from actual network devices included in the physical network layer of the real world (e.g., mobile access network, etc.).

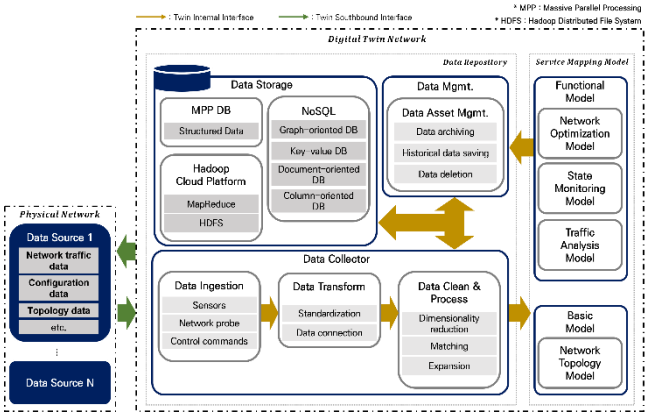


Fig. 3. Example of the network data lifecycle process.

The data used internally in DTN is collected and processed by a data collector in the data repository. This follows three main processes. First, it goes through data acquisition from sensors or analyzers, followed by a data transformation step for standardized representation of heterogeneous devices. Next, it goes through a data cleaning and processing step, which involves dimensional reduction or merging of necessary data, depending on the task within the DTN.

The data collected and processed in this way is made available as a service mapping model or stored in data storage according to the structured information of the data. When storing DTN data in a data storage, the data type and structural characteristics should be considered. Most of the DTN modeling data is structured data, so the main Database(DB) in data storage can be built based on Massive Parallel Processing(MPP) database. For unstructured and semi-structured data, it can be stored and processed based on the Hadoop platform, and parallel processing of tasks can be done using MapReduce and file storage using Hadoop Distributed File System(HDFS). NoSQL databases can also be used to store unstructured and semi-structured data, and they can be designed and used with data types in mind, such as graph-oriented DB and key-value DB.

The stored data can be reorganized by data management and provided as a service mapping model, and the execution results of the model within the service mapping model can be stored or deleted again by data management. Through this process, DTN network data can be managed.

## V. CONCLUSION

This paper proposes a data pipeline for network data management in DTN, along with the types of data available in DTN. The proposed method is designed to enable continuous operation and management of the communication network in DTN by processing, storing, and reconstructing data collected from physical entities located in the actual network. In addition, the pipeline includes multiple storage options based on structured information according to the data type, and also includes explanations for processing the results derived from the service mapping model. This pipeline, designed with these data types in mind, will help to effectively perform network data management within the DTN in the future. In addition, we intend to conduct research on artificial intelligence models based on DTN scenarios that utilize the data managed by this structure in the future.

## ACKNOWLEDGMENT

This work was supported by Electronics and Telecommunications Research Institute(ETRI) grant funded by ICT R&D program of MSIT/IITP[2019-0-00260, Hyper-Connected Common Networking Service Research Infrastructure Testbed] and supported by the MSIT(Ministry of Science and ICT), Korea, under the Innovative Human Resource Development for Local Intellectualization support program(IITP-2023-RS-2022-00156287) supervised by the IITP(Institute for Information & communications Technology Planning & Evaluation).

## REFERENCES

[1] CLEMM, Alexander; ZHANI, Mohamed Faten; BOUTABA, Raouf. Network management 2030: Operations and control of network 2030 services. Journal of Network and Systems Management, 2020, 28.4: 721-750.

[2] Tao, Sun, et al. Digital twin network (DTN): concepts, architecture, and key technologies. Acta Automatica Sinica, 2021, 47.3: 569-582.

[3] Gartner. Rethink Network Monitoring for a Cloud Era; Gartner: Singapore, 2018.

[4] Jérôme François, Laurent Ciavaglia, et al. Network Management (nmrg), [online] available: https://datatracker.ietf.org/rg/nmrg/about/, Oct. 1st, 2023.

[5] Zhou, C.; Yang, H.; Duan, X.; Lopez, D.; Pastor, A.; Wu, Q.; Boucadair, M.; Jacquenet, C. Digital Twin Network: Concepts and Reference Architecture (draft-irtf-nmrg-network-digital-twin-arch-03). In IRTF Internet-Draft; Internet Engineering Task Force: Fremont, CA, USA, 2023.

[6] Li, M.; Zhou, C.; Chen, D. Data Generation and Optimization for Digital Twin Network Performance Modeling (draft-li-nmrg-dtn-data-generation-optimization-00). In IRTF; Internet Engineering Task Force: Fremont, CA, USA, 2023.

[7] Shin, H., Oh, S., Isah, A., Aliyu, I., Park, J., & Kim, J. (2023). Network Traffic Prediction Model in a Data-Driven Digital Twin Network Architecture. Electronics, 12(18), 3957.

[8] Yang, H.; Lü, P.; Sun, T.; Lu, L.; Zhou, C. Multi-source Heterogeneous Data Processing Technology for Digital Twin Network. In Proceedings of the 2022 IEEE 22nd International Conference on Communication Technology (ICCT), Nanjing, China, 11–14 November 2022; pp. 1829–1834.

[9] Lo, Wai Weng, et al. "E-graphsage: A graph neural network based intrusion detection system for iot." NOMS 2022-2022 IEEE/IFIP Network Operations and Management Symposium. IEEE, 2022.

[10] Zhou, Jiawei, et al. "Automating botnet detection with graph neural networks." arXiv preprint arXiv:2003.06344 (2020).

[11] Rusek, Krzysztof, et al. "Routenet: Leveraging graph neural networks for network modeling and optimization in sdn." IEEE Journal on Selected Areas in Communications 38.10 (2020): 2260-2270.

[12] Mallick, Tanwi, et al. "Dynamic graph neural network for traffic forecasting in wide area networks." 2020 IEEE International Conference on Big Data (Big Data). IEEE, 2020.

[13] Wang, Mowei, et al. "xnet: Improving expressiveness and granularity for network modeling with graph neural networks." IEEE INFOCOM 2022-IEEE Conference on Computer Communications. IEEE, 2022.

[14] Liu, Wai-xi, et al. "DRL-R: Deep reinforcement learning approach for intelligent routing in software-defined data-center networks." Journal of Network and Computer Applications 177 (2021): 102865.