Data Plane enabled datasets for AI-native network

Lu Lu
China Mobile

16 July 2024

Contents

1 AI/ML datasets

2 Data Plane enabled datasets

Perspectives on datasets

Data is the key element of AI. One of the visions for AI datasets is to provide high-quality and diversified data resources to support the training and optimization of AI algorithms.

Data Validity

The data value is consistent with the valid value or valid reference range of the definition.

Data Consistency

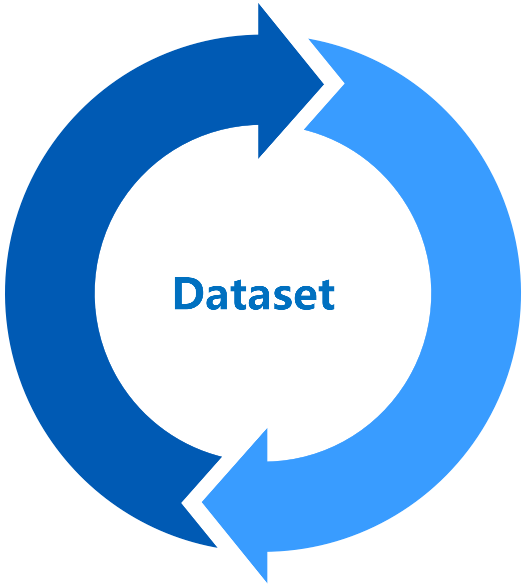
The attributes of data stored in different distributed nodes are consistent.

Data Uniqueness

The elements in the datasets do not appear repeatedly.

Data Timeliness

The dataset can reflect current or recent real-time information or status for AI.



Data Completeness

The data contains all necessary information without any omissions or missing parts.

Data Integrity

Both historical business data and timely updated data after the model goes online are required.

Data Rationality

The comparison with benchmark data helps to determine whether the distribution, and modality of the data are reasonable.

Data Accuracy

The data must accurately reflect the facts and must not contain any errors, false or misleading information.

Challenges of network datasets

For network AI, the data standardization level and data sources are more diverse and complex. It is challenging to design efficient mechanisms for AI data collection, transmission, processing, and storage to meet the demands of the network.

Missing Data

Inconsistent Data

Incomplete Data

Inaccurate Data

Non-standard Data

- How to subscribe/collect data on demand while avoiding duplicate collection in order to minimize network overhead?
- How to achieve real-time data collection at the UE level, business level, or with different time granularities?
- How to support cross-domain data collection covering environments, user behaviors, services experiences, etc.?
- How to address the issues of partial data missingness and imbalanced sample sets?
- How can we achieve automatic annotation and association of data?
- How to standardize common network data field types, such as timestamps?

Innovation direction for network datasets

To address the challenge of data complexity and data missing, data governance and data exposure are the main research direction.

Data Governance: Enhancing Data Standardization and Quality

1. Collaborate to jointly develop **data collection standards** for network.
2. On-demand **dynamic data collection granularity** solutions.
3. **Real-time validation capabilities** to promote timely quality improvements.
4. **Closed-loop auditing capabilities** to ensure data accuracy and reliability.
5. **Unified processing capabilities** to achieve compatibility with various types of data.

Data Exposure: Promoting the efficient utilization and sharing of data in a secure manner

- QoE Perception Optimization
- Graph Topology Optimization
- Multi-Indicator Spatiotemporal Sequence Prediction
- Root Cause Analysis
- Network Resource Image Recognition
- Cloud Network Anomaly Detection
- Channel Measurement Dataset (Channel Compression)
- Intelligent RAN Slicing
- Network AIScheduling Technology Dataset
- Cell freeScenario-based Wireless Resource Scheduling
- Location
- Beam Management
- ……

Datasets practices of China Mobile

中国移动 China Mobile

China Mobile has launched 15 premium AI datasets, providing billion-scale core resources to enable network and AI capabilities

Currently, this series of datasets are built to realize the following capabilities through consensus-based data governance technology, and aims to further expose to the industry for collaborative datasets sharing.

Perception | **Diagnosis** | **Prediction** | **Make decision** | **General AI** | **Large model**

Capability Areas

Self-built dataset

CSI measurement data 7500+
Intelligent NF: Prediction/Optimization

Container network metrics data 1060,000+
Network operation: Diagnosis/Maintenance

Long-term network traffic prediction 105,000+
Network operation: Prediction/Operation/Optimization

Uplink interference identification 15,000+
Network optimization: Perception/Diagnosis/Operation

Industry partners co-build dataset

AI air interface channel simulation data 100 million+
Intelligent NF: Prediction/Optimization

Security event situational awareness 15,000+
Network operation: Perception/Diagnosis/Operation

Intelligent network traffic classification 2900+
Intelligent Services: Perception/Diagnosis/Operation

Intelligent network routing selection 100+
Intelligent NF: Decision-Making/Operation

Contents

1 AI/ML datasets

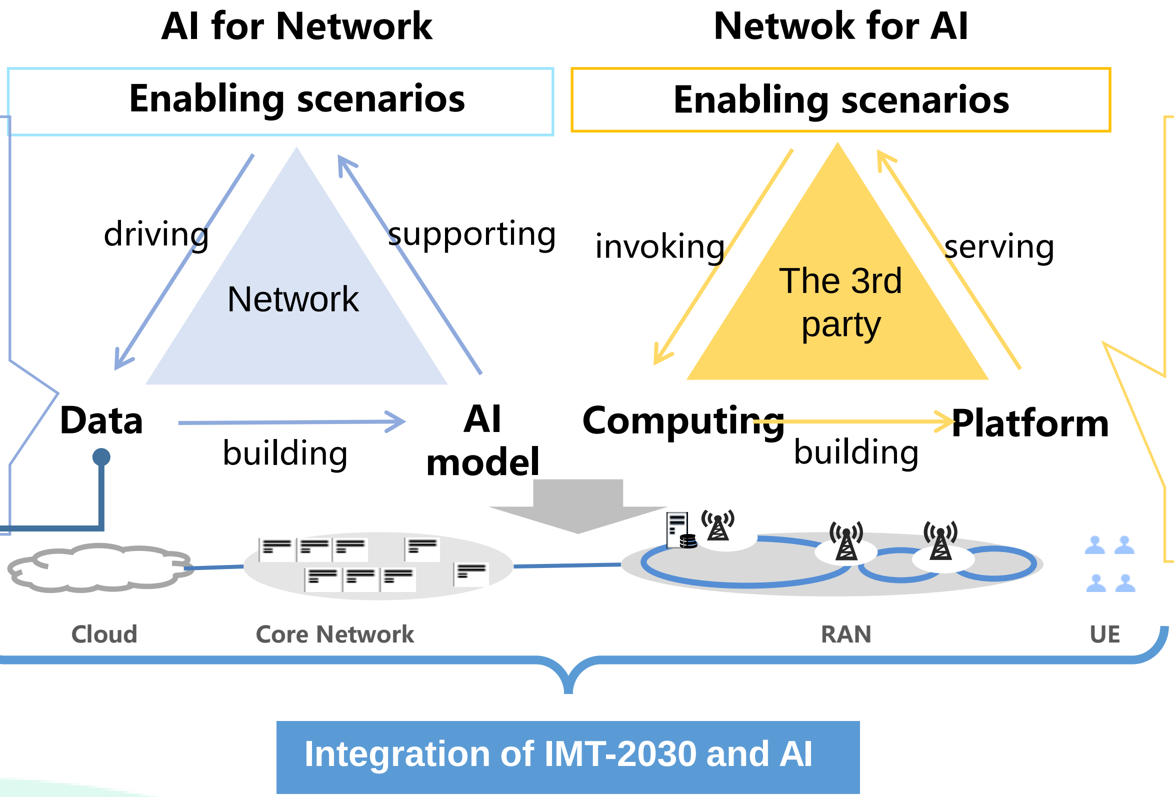
2 Data Plane enabled datasets

AI-native network

To enable the "Integrated AI and Communication" usage scenario for IMT-2030, the end-to-end AI-native network architecture is required to make the data, computing and algorithm of AI as well as the network connection become the foundation of network.

By building the AI-native network environment and capability, it can improve the network operation and maintenance performance (AI for network) and enable the full lifecycle of AI (network for AI).

Driven by scenarios, it focuses on solving the problems of **network data governance** and **network AI model construction** to improve the efficiency of network O&M



Driven by scenarios, it focuses on solving the problem of ubiquitous **computing** and **capability platform** construction, supporting data /AI model transmission and providing AI services

Intrinsic data needs to be collected, processed and organized more effectively to build the datasets.

Why data plane is needed

中国移动 China Mobile

To support the datasets for AI-native network, data lifecycle management is needed

The current network is only used as a data transmission pipeline and cannot meet the collection, transmission, processing and storage requirements of intrinsic datasets.

Build an independent data plane for the whole lifecycle of data management for building intrinsic datasets.

Requirement and Challenge

Requirements	data collecton	data transmission	data processing	data storage
challenge	• Real-time; • Fine grained; • Non-per UE	• Large amount of data; • High concurrency; • A variety of QoS requirements	• De-privacy processing; • Data processing; • Data/model encapsulation	• Training/reasoning etc require a lot of data storage; • Unstructured data storage such as AI models; • Fast index
Existing user plane	Not supported	Partial support	Not supported	Not supported
Existing control plane	Partial support	Partial support	Partial support	Partial support

Design ideas

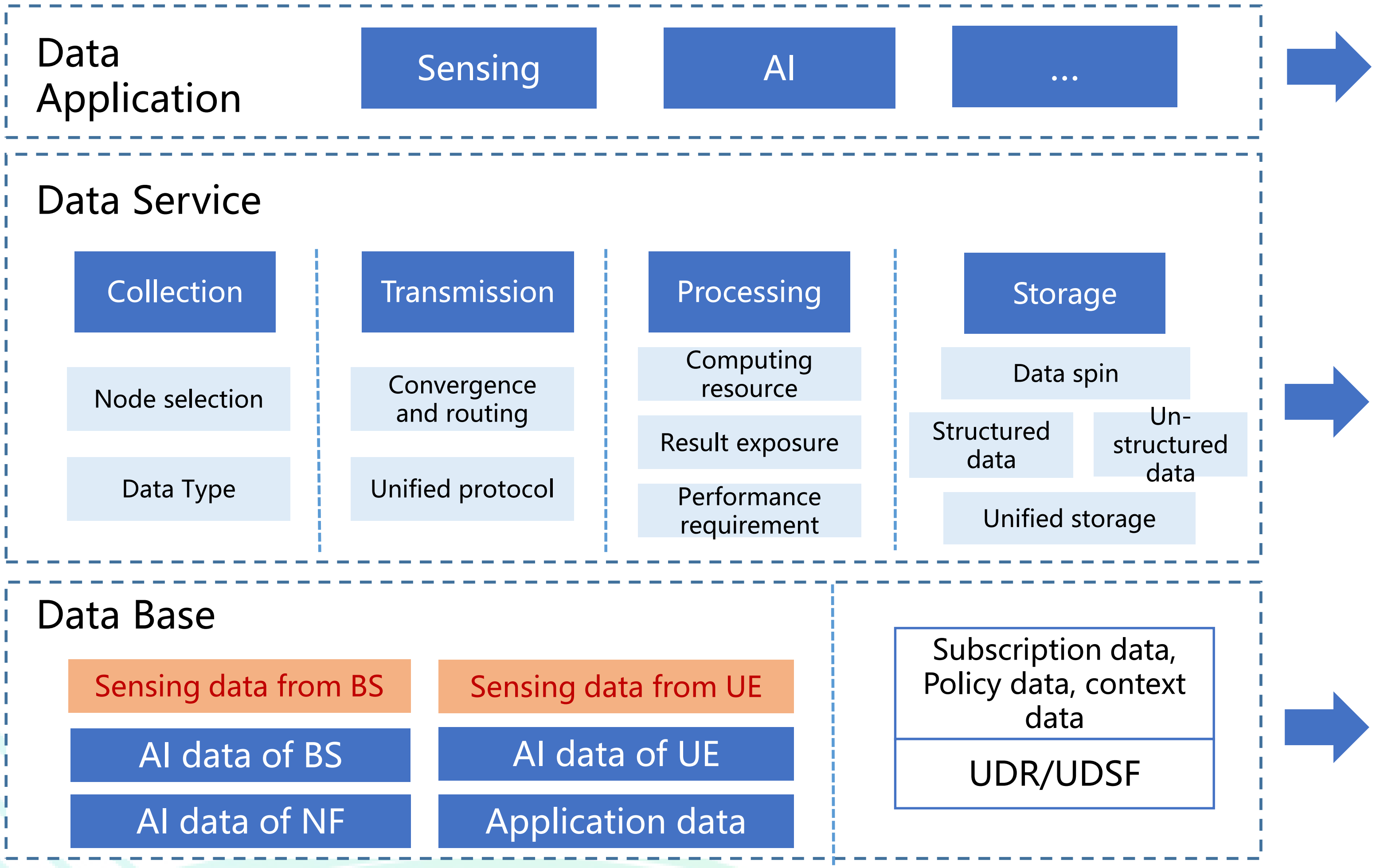
①A new function set needs to be designed
②The performance and mechanism of the new feature set are different
③Unified control combined with communication characteristics

New data plane

Enhanced control plane

Data plane is designed to support the whole process of data production and consumption

Logical function design for data plane



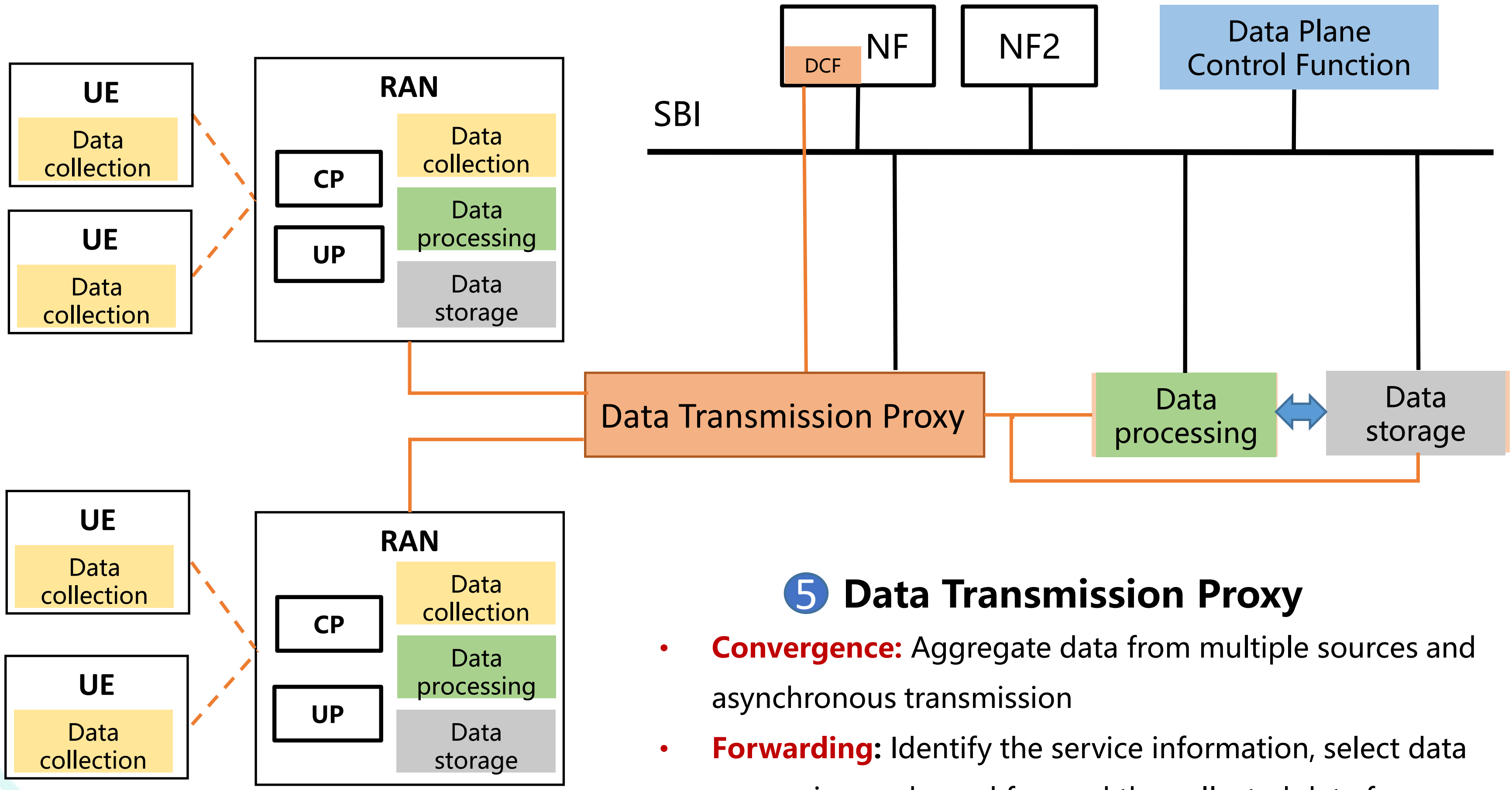
new applicaitons supported by data base and data services, e.g. sensing, AI, digital twin, etc.

data plane includes basic data services, e.g. collection, transmission, processing, storage, and supports the whole processes of data management.

data collected from UE, base stations, NFs, and applicaitons, as well as traditional user communication data

Y.IMT2020-DDP: "Future networks including IMT-2020: requirements and framework of distributed data plane"

Architecture design for data plane



5 Data Transmission Proxy

- **Convergence:** Aggregate data from multiple sources and asynchronous transmission
- **Forwarding:** Identify the service information, select data processing node, and forward the collected data from base stations, UEs, and NFs of distributed CN
- **Network Topology Aggregation:** Avoid multi-path transmission tunnel establishment to achieve efficient data transmission.

1 Data Plane Control Function

- Selection and authentication of data collection node
- Construction of data transmission path

2 Data Collection Function

- Data collection from multiple sources and multiple nodes

3 Data Processing Function

- Data convergence of multiple sources

4

- On-demand network internal processing

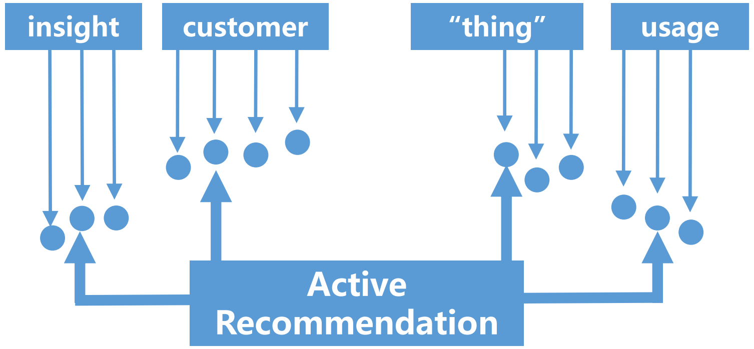
Data Storage Function

- Unified data storage and management

Key technologies for data plane

Data fabric

- Data fabric technology can enhance data integration and data operation supply capabilities across data centers, domains, and vendors, facilitating unified data management and efficient data collection.

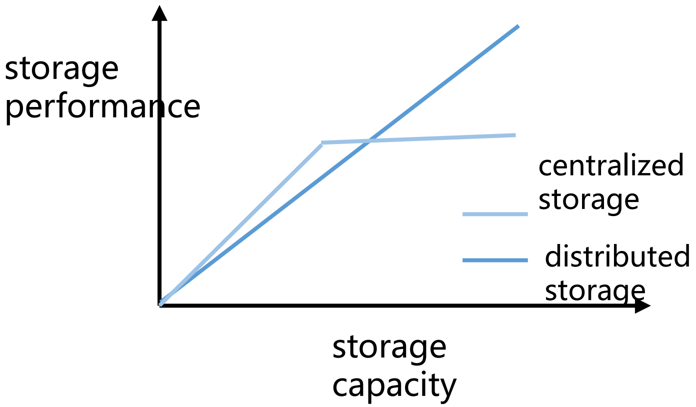


Data privacy protection

- The security and privacy requirements of data plane are higher.
- Multi-point cooperative learning methods such as federated learning and homomorphic encryption techniques can be considered

Distributed data transmission and storage

- Distributed network is becoming the solution to solve large-scale data processing and high concurrent access
- Distributed data transmission and storage is the key technology for efficient collection and trusted sharing of distributed data



6G Architecture Design: from Overall, Logical and Networking Perspective. IEEE Commun. Mag. 61(7): 158-164 (2023)

Summary and Prospect

The quality and diversity of network datasets directly impact the availability and effectiveness of network AI, playing a pivotal role in advancing and implementing network AI technologies.

AI-native networks reqiures intrinsic datasets, and the data plane can promote the implementation of intrinsic datasets. The design of data plane is expected to better enable datasets to support the realization of AI-native capabilities in IMT-2030.

China Mobile would like to collaborate with industry partners to construct high-quality network datasets, to facilitate innovation in the AI-native network for IMT-2030 based on the findings elaborated in ITU-T SG13.