

Cloud computing strategic alliance technical standards

HB/T-2020-0001

The Technology Standards of Data Resource Distribution Map and In-situ Virtual Data Center

Drafted by: Tongji University, Donghua University

Release date: September 1, 2020



Preface

"Data Resource Distribution Map and In-Situ Virtual Data Center" consists of the following 7 parts:

- _____ Part 1: Scope;
- _____ Part 2: Normative references;
- _____ Part 3: Overview;
- _____ Part 4: Data resource distribution map;
- _____ Part 5: In-situ virtual data center system framework;
- _____ Part 6: The core construction process of the in-situ virtual data center;
- _____ Part 7: The structure of network data explorers and the generation of resource distribution maps.

This standard was drafted in accordance with the rules given in GB/T 1.1-2009.

The translation method used in this standard is equivalent to the national standard GB/T 25000.51-2016 "System and Software Engineering System and Software Quality Requirements and Evaluation (SQuaRE) Part 51: Quality Requirements and Test Rules for Ready-to-Use Software Products (RUSP)".

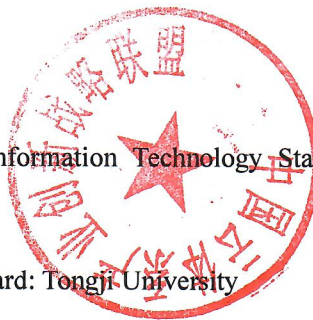
This standard was proposed by Tongji University.

This standard is under the jurisdiction of the Information Technology Standardization Technical Committee (SAC/TC 180).

The organization responsible for drafting this standard: Tongji University

Participated in the drafting of this standard: Donghua University

The main drafters of this standard: Jiang Changjun, Zhang Zhaohui, Ding Zhijun, Yu Jian, Yan Chungang, Zhang Yaying



Introduction

The Internet is a dynamic, open, and shared system. Therefore, Internet data is not only huge in scale, numerous in sources, diverse in types, and highly dynamic. This makes it difficult for data analysts or data users to clearly understand the Internet big data, that is, where the data is needed, how much data, and what the data components are. Data demanders often take the most comprehensive approach to collect data and analyze to fully utilize data. This causes a large amount of redundant and useless data, which causes a waste of resources. Therefore, how to provide data demanders with effective data sources and their basic data characteristics without collecting and storing a large amount of useless data has become an urgent problem for big data.

Existing big data centers mainly adopt full data collecting, analyzing, and processing methods. There is blindness in data acquisition and disorder in resource utilization, which greatly wastes computing resources, storage resources, and energy. The in-situ virtual data center is designed to meet the needs of data collection, analysis, service, and data security and credible management and control. It does not collect and store the full amount of data from the target source itself, but generates data resource distribution through the exploration of Internet data resources. It also provides data guidance services for data analysts to solve the problem of blindness and disorder in Internet data acquisition, analysis and utilization.

This standard mainly introduces the data resource distribution map and the construction technology system of the in-situ virtual data center. Given the corresponding specifications, the generation of data resource distribution maps and the construction of in-situ virtual data centers can be realized.

The Technology Standards of Data Resource Distribution Map and In-situ Virtual Data Center

1 Scope

This standard specifies the data resource distribution map and the architecture of the in-situ virtual data center, and provides a unified name specification and definition description for it. It also provides a reference for the development of other standards of the data resource distribution map and the in-situ virtual data center.

This standard establishes the general principles of all data resource distribution maps and in-situ virtual data centers, as well as products and systems design, research and development, release, management, and maintenance. It will provide reference for data resource distribution maps and in-situ virtual data center construction indicator specifications.

2 Normative references

The following documents are indispensable for the application of this document. For dated reference documents, only the dated version applies to this document. For undated references, the latest version (including all amendments) applies to this document.

GB/T 1.1-2009 Standardization Guidelines

ISO/IEC 23006-4-2013 Information Technology Multimedia Service Platform Technology
Part 4: Basic Service

ISO/IEC 23006-1-2013 Information Technology Multimedia Service Platform Technology
Part 1: Architecture

AS 3965-1991 Information Technology Open System and Interconnection Public
Management Information Service Definition

ISO 5807-1985 Information processing data flow, program flow chart, system flow chart,
program network diagram, system resource diagram documentation symbols and conventions

GB/T 16680-1996 Software Document Management Guide

GB/T 25000.51-2016 System and Software Engineering System and Software Quality

Requirements and Evaluation (SQuaRE) Part 51: Quality Requirements and Test Rules for Ready-to-Use Software Products (RUSP)

GB 33473-2016 General technical requirements for HI interface of instant messaging service

JR/T 0011-2004 Bank Centralized Data Center Specification

3 Overview

Many existing big data centers mainly obtain Internet data through mass collection, organize and process the data, and then provide application support to customers. Specifically, there are two main ways to obtain data. One is to collect URL information and then collect data by crawler-like network robots. The other is to call the API interface to implement the collection of the internal database of the data source according to the method in the DB API protocol. Regardless of the data collection method, as long as the Internet data is collected in large quantities, a large amount of redundant and low information value data is often obtained. This not only consumes huge amount of resources and service performance of the data source, but also consumes a large amount of computing, storage, and network resources of the data collector. Data demanders often take as full data as possible to collect data and analyze and use data, making it difficult for data analysts or data users to clearly understand Internet big data, that is, where the data is needed, how much data is there, and what are the data components. The problem is not clear. The data resource distribution map of this standard mainly reflects the basic characteristics of network big data, and provides data analysts with the basic situation of network big data sources.

This standard proposes the in-situ virtual data center architecture and construction methods, aiming to meet the data collection and analysis needs of the data center, and form relatively orderly and effective acquisition and analysis of Internet data resources, Internet data access, and Internet data content so as to provide data resource distribution guidance services for data centers, and provide data access protocols for data providers. Its characteristics include the following aspects: (1) Compared with the existing data center, the in-situ virtual data center is a lightweight data center. It does not collect and store the full amount of data from the target source by itself, but is connected to the Internet. Data exploration, virtualize Internet data resources, and form a data resource distribution map. (2) The core of the in-situ virtual data center is the network data explorer and data resource distribution map. The information volume and value density of the website are explored through the network data explorer, and the data size, value density and distribution of the network site are obtained by using the sample estimation method. Then the data resource distribution map is generated. According to the data

resource distribution map and the data resource acquisition guidance service, the approximate distribute on and value information of the data source and its Internet data resources will be provided to data demand method.

4 The architecture of data resource distribution map

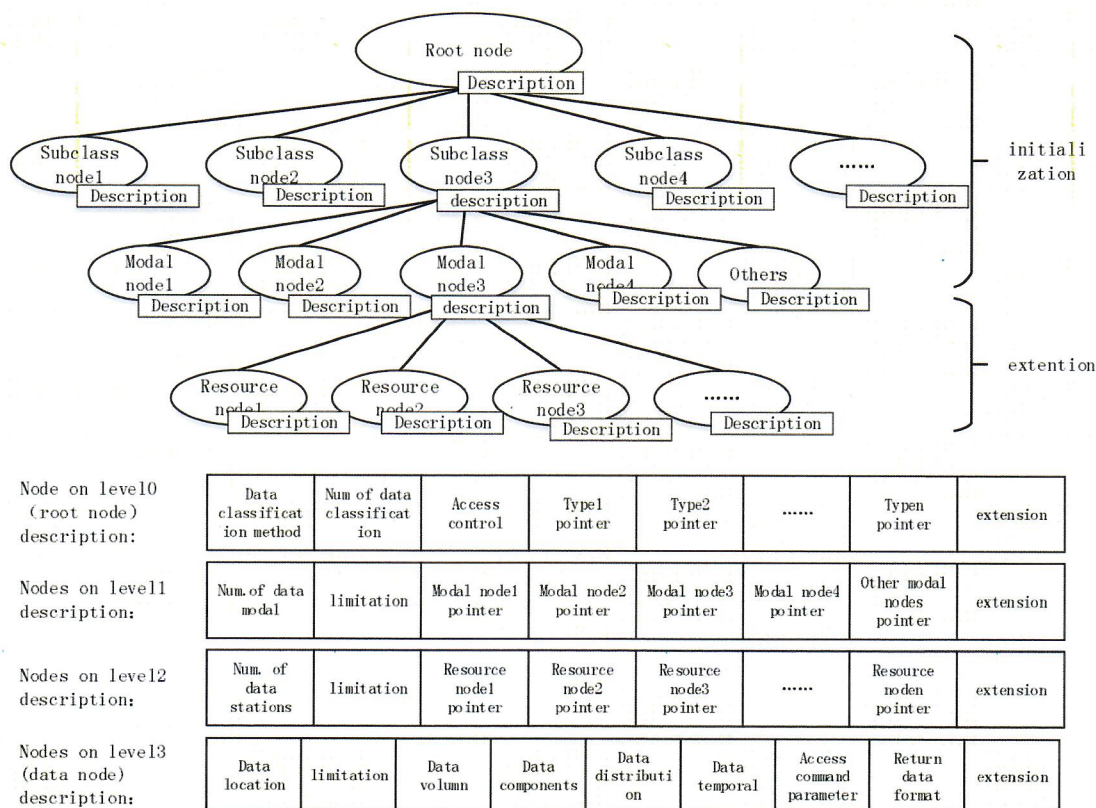


Figure 1 Data resource distribution map structure

The structure of the data resource distribution map is shown in Figure 1, which mainly includes layer 0 nodes (root nodes), layer 1 nodes, layer 2 nodes, and layer 3 nodes (data nodes). Among them, layer 0 nodes (root nodes), layer 1 nodes, and layer 2 nodes are initialization layer nodes, layer 3 nodes (data nodes) are expansion layer nodes, and layer 4 nodes form a tree structure. The specific structure and its structure are described as follows:

- (1) The 0th layer node (root node) mainly includes data classification method, data classification number, access restriction, category 1 pointer, category 2 pointer..., category n pointer and extension items, etc. Among them, the data classification

method item record is used for the data classification model or method; the category pointer is used to point to the category node, that is, each child node of the root node is a category, and the extension item is used for information expansion.

- (2) The first layer nodes mainly include descriptions of data modal numbers, restricted commands, text pointers, image pointers, video pointers, voice pointers, other pointers, and extension items. The number of data modalities refers to the number of classifications of data modalities. Generally, it refers to five types of data: text, image, video, voice, and others; text pointers, image pointers, video pointers, voice pointers, and other types of pointers are the link pointer to the child node, whose child node is a node of a certain data modal.
- (3) The second layer nodes mainly include descriptions of the number of data sites, limit commands, site 1 pointer, site 2 pointer, ..., site m pointer and extension items, etc. The number of data sites refers to the total number of data source sites in a certain type of data mode. The number also indicates the number of nodes of its children; the site pointer records each of its child nodes.

The third layer nodes are data nodes, which mainly include data location, limit commands, data volume, data components, data distribution, data sequence, access commands and parameters, return data format, extended items and other descriptions. The data location records the site location of the data source; the restricted command is the restricted access description for accessing the data source; the data volume is the data quantity of the site, provided by the data provider (it can also be empty); the data component indicates the constituent elements of the data ; Data distribution is the basic characteristics of the data and its distribution; the time sequence of the data indicates whether the data has a time series relationship; the access command and parameter record the command and its parameters to access the data source (it also can be empty); return data format refers to the format of the acquired data.

5 The architecture of the in-situ virtual data center system

The architecture of the in-situ virtual data center system is shown in Figure 2.

data resource distribution map generating unit is used to generate a data resource distribution map based on the attribute information of the Internet data and the access restrictions in the data sampling guide tree.

(2) Internet virtual resource library. It is used to store the data resource distribution map and the sample data collected by the Internet data explorer, mainly including the data resource distribution map and the exploration sample library. The data resource distribution map is the core data structure component of the in-situ virtual data center. It reflects the overall distribution of Internet data, including data location, data volume, data characteristics etc. It is a guide information table for large-scale data collection, as shown in the figure 1. The exploration sample library is used to store small sample data collected by the network data explorer.

(3) Data resource distribution map management subsystem. It is a management system for storing, accessing, and updating data resource distribution maps. Wherein, the data resource distribution map is stored in a relational or non-relational database; the data resource distribution map is accessed according to a tree structure. The core of the data resource distribution map management in this standard is the dynamic update method of the data resource distribution map, which will ensure that the Internet virtual resource library is up to date.

(4) Data resource acquisition guidance service subsystem. It is based on the resource distribution map to provide data collection and mining guidance services to the data demander to ensure that data demand users can efficiently and orderly collect and mine Internet data and further analysis.

(5) Data protocol generation and management subsystem. It is based on the data access protocol provided by the Internet data provider and the data site map to generate a unified data access protocol file, including Web data access protocol, Internet internal database access protocol, etc., and can provide management functions for these protocols, including protocol Publish, update, etc.

(6) Data security and trustworthy management subsystem. It is used to perform data security management on virtual data resources in the Internet virtual resource library, and mainly access management of virtual data resources, including management of data privacy protection, data access authority, etc.

6 The core construction process of the in-situ virtual data center

The core structure of the in-situ virtual data center mainly includes network data explorer, resource distribution map structure and management. Proceed as follows:

Step 1: Build a network data explorer based on the data protocol and Internet big data. The detailed process is as follows:

- a) According to the data protocol, the explorer module is divided into two types: Web page data and internal database API for sampling guidance;
- b) Analyze the results of data sampling and guidance, establish a guidance tree for Web page data, and establish a guidance table for the internal database;

Step 2: Construct an Internet virtual resource database based on the data results of the network data explorer's exploration, including the exploration sample database and the data resource distribution map;

- a) According to the results of data sampling guidance, the explorer module conducts data sampling estimation on Internet big data into two types: Web page data and internal database, and estimates the value of data including data volume, data composition, data distribution, etc., and builds exploration sample library;
- b) Generate a data resource distribution map based on the data sampling guide tree/guide table and the results of data sampling estimation.

Step 3: Manage the data resource distribution map according to the results of network data exploration and the data resource distribution map. The detailed process is as follows:

- a) Provide data resource acquisition guidance services according to the data resource distribution map;
- b) According to the data protocol, generate and manage the data protocol of the in-situ virtual data center according to different data providers;
- c) Maintenance and management of Internet virtual resource database, including exploration sample database and data resource distribution map.

7 The structure of network data explorer and the generation of resource distribution map

The network data explorer is mainly composed of four parts: data sampling guide module, data sampling guide tree/guide table, data sampling estimation module, and data resource distribution map generation module. As shown in Figure 3, its working principle is:

- (1) The data sampling guide module mainly generates data sampling guide information according to the relevant data access restrictions of the data provider. It is mainly divided into two types of guidance: one is Web page data sampling guidance, and the other is internal database API sampling guidance. Web page data sampling guide is mainly to read data crawling protocol files and site map files in the Internet, and read part of the data according to a certain strategy to generate a data sampling guide tree. The data sampling guide tree records the accessible data site resources and their access permissions. The internal database API sampling guide is mainly to generate a data sampling guide tree by reading the access method and access restricted standard access file provided by the data provider; if the standard access restriction file is not provided, the standard access file is manually configured and then generated Data sampling guide tree.
- (2) The data sampling guide tree/guide table is the data structure of the data sampling guide information generated by the data sampling guide module.
 - a) The data sampling guide tree refers to the guide information for sampling Web information, which is generated by the Web page data sampling guide module, as shown in Figure 4. The Web data sampling guide tree is mainly a tree structure. The root node is the root directory node of the website, and the sub-nodes are the sub-directory nodes of the sub-site. The description items of each node include the data location (the site location where the data is located), data Modal (text, image, video, voice, etc.), data explorer name, data access restriction commands, data timing characteristics, access commands, command parameters, returned data format (page or Jason and other data formats), extension items (Used for the extended description of other Web forms of data).

- b) The data sampling guide table refers to the data sampling guide information table for accessing the internal database through the API interface on the Internet, as shown in Table 1. Mainly include data location (location of the site where the data is located), data modal, data explorer name, access prohibited/restricted items, API call function table (including parameters and return values) description, data sequence, data distribution, Whether the data is online/offline, and extended items.
- (3) Data sampling estimation module root Web data sampling guide tree and internal database API sampling guide table, according to a certain strategy (interval sampling or point sampling strategy) to capture a certain amount of data and store it in the Internet virtual resource library; at the same time Internet Web Data sampling estimation and/or internal database application programming interface sampling estimation, estimation of data category, data modal, data volume, data composition, data distribution, etc.
- (4) The data resource distribution map generation module generates a data resource distribution map based on the analysis results of data sampling and the access restrictions in the data sampling guide tree.

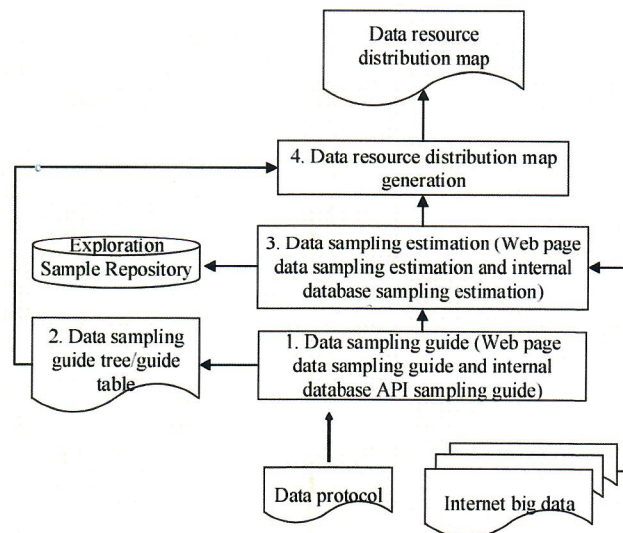


Figure 3 The structure of the network data explorer

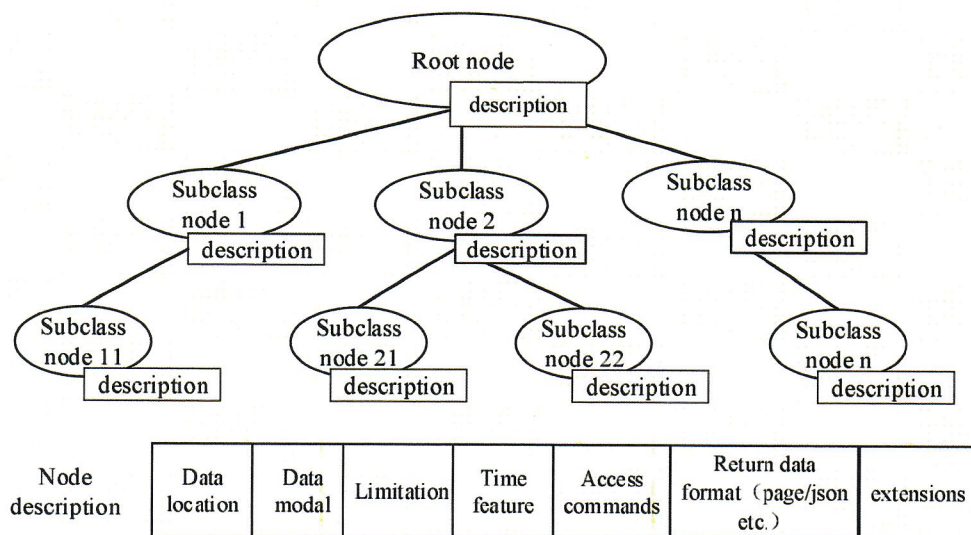


Figure 4 The constitution model of Web-GuideTree

Table 1 The internal database API sampling guide information table API-GuideList

Resource location	Data modal	Data Explorer name	Access prohibited/restricted items	API call function table (including parameters, return value) description	Data timing	Distribution of data	Whether the data is online/offline	Extension

Reference

- [1] GB/T 1.1-2009 Standardization Guidelines
- [2] GB 4943.1-2011 Information Technology Equipment Safety Part 1: General Requirements
- [3] ISO/IEC 23006-4-2013 Information Technology Multimedia Service Platform Technology Part 4: Basic Service
- [4] ISO/IEC 23006-1-2013 Information Technology Multimedia Service Platform Technology Part 1: Architecture
- [5] ISO/IEC 23006-5-2013 Information technology Multimedia service platform technology Part 5: Service aggregation
- [6] GB/T 29746-2013 Real-time traffic information service data structure
- [7] JR/T 0011-2004 Bank Centralized Data Center Specification
- [8] GB 4943.23-2012/IEC 60950-23:2005 Information Technology Equipment Safety Part 23: Large Data Storage Equipment
- [9] ISO 5807-1985 Information Processing Data Flow Diagram, Program Flow Diagram, System Flow Diagram, Program Network Diagram and System Resource Diagram Documentation Symbols and Conventions
- [10] GB/T 16680-1996 Software Document Management Guide
- [11] GB 33473-2016 General technical requirements for HI interface of instant messaging service
- [12] GB/T 32399-2015 Information Technology Cloud Computing Reference Framework
- [13] GB/T 25000.51-2016 System and Software Engineering System and Software Quality Requirements and Evaluation (SQuaRE) Part 51: Quality Requirements and Test Rules for Ready to Use Software Products (RUSP)
- [14] GB/T 32423-2015 System and Software Engineering Verification and Confirmation
- [15] GB/T 25067-2016/ISO/IEC 27006:2011 Information technology Security technology Information security management system audit and certification body requirements