中国云体系产业创新战略联盟
China Cloud System Pioneer Strategic Alliance

云计算战略联盟技术标准

HB/T-2020-0001

数据资源分布图与原位虚拟数据中心技术标准

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

编制单位：同济大学、东华大学

发布时间：2020-09-01

前　言

《数据资源分布图与原位虚拟数据中心》由以下 7 部分组成：

_____第 1 部分：范围；

_____第 2 部分：规范性引用文件；

_____第 3 部分: 概述；

_____第 4 部分：数据资源分布图；

_____第 5 部分：原位虚拟数据中心系统的框架；

_____第 6 部分：原位虚拟数据中心的核心构造流程；

_____第 7 部分：网络数据勘探器的构造及资源分布图的生成。

本标准按照 GB/T 1.1-2009 给出的规则起草。

本标准使用翻译法等同采用国家标准 GB/T 25000.51-2016《系统与软件工程　系统与软件质量要求和评价(SQuaRE)　第 51 部分：就绪可用软件产品（RUSP）的质量要求和测试细则》。

本标准由同济大学提出。

本标准由信息技术标准化技术委员会（SAC/TC 180）归口。

本标准负责起草单位：同济大学

本标准参加起草单位：东华大学

本标准主要起草人：蒋昌俊、章昭辉、丁志军、喻剑、闫春钢、张亚英

引　言

互联网是一个动态的、开放的、共享的系统，因此，互联网数据不但规模巨大、来源众多、种类多样，而且动态性强。这使得数据分析者或数据利用方难以清晰地认知互联网大数据，即所需的数据在哪里、数据有多少、数据成分是什么等问题不清楚，数据需求者往往是采取尽可能全量方式去采集数据并分析利用数据。这造成了数据采集方会采集大量的冗余无用的数据，从而造成资源的浪费。因此，如何为数据需求者提供有效的数据源及其基本数据特征，避免数据需求者采集存储大量无用数据而造成资源浪费，成为大数据亟待解决的问题。

现有的大数据中心主要采用全量的数据采集、分析、处理等方法，存在数据获取的盲目性、资源利用的无序性，极大地浪费了各种计算资源、存储资源以及能源。原位虚拟数据中心旨在面向数据中心的数据采集、分析、服务以及数据安全可信管控的需求，其自身并不采集和存储目标源全量数据，而是通过互联网数据资源的勘探生成数据资源分布图，并为数据分析者提供数据制导服务，以便于解决互联网数据获取、分析和利用的盲目性和无序性问题。

本标准主要介绍了数据资源分布图和原位虚拟数据中心的建构技术体系。通过给定相应的规范标准，可以实现数据资源分布图的生成和原位虚拟数据中心的建构。

数据资源分布图与原位虚拟数据中心技术标准

1 范围

本标准规定了数据资源分布图与原位虚拟数据中心建构体系，对其进行了统一的名称规范和定义说明，并为数据资源分布图与原位虚拟数据中心其它各项标准的编制提供参照。

本标准确立了所有数据资源分布图与原位虚拟数据中心及其设计、研制、发行、管理、维护的产品、系统等的一般原则，为数据资源分布图与原位虚拟数据中心构建提供参照性指标规范。

2 规范性引用文件

下列文件对于本文件的应用必不可少的。凡是注日期的引用文件，仅所注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 1.1-2009 标准化工作导则

ISO/IEC 23006-4-2013 信息技术 多媒体服务平台技术 第 4 部分:基本服务

ISO/IEC 23006-1-2013 信息技术 多媒体服务平台技术 第 1 部分:架构

AS 3965-1991 信息技术 开放系统和互联 公共管理信息服务定义

ISO 5807-1985 信息处理 数据流程、程序流程图、系统流程图、程序网络图、系统资源图的文件编制符号及约定

GB/T 16680-1996 软件文档管理指南

GB/T 25000.51-2016 系统与软件工程 系统与软件质量要求和评价（SQuaRE） 第 51 部分：就绪可用软件产品（RUSP）的质量要求和测试细则

GB 33473-2016 即时通信业务 HI 接口总体技术要求

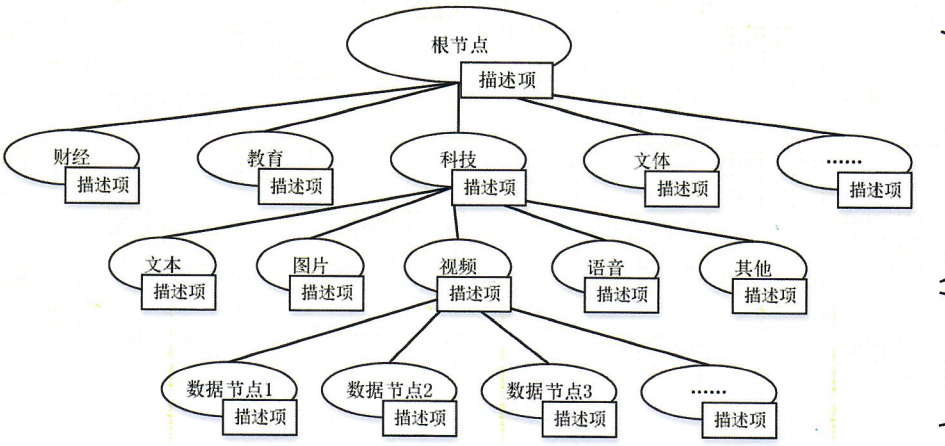
JR/T 0011-2004 银行集中式数据中心规范

3 概述

现有很多的大数据中心主要通过大批量采集方式得到互联网数据，并对数据进行整理和加工，进而对客户提供应用支持。具体来说，主要有两种方式获取数据。一种是通过爬虫类的网络机器人的方式采集 URL 信息进而采集数据。另一种是根据 DB API 协议中的方法，调用 API 接口实现数据源内部数据库的采集。不管是哪种数据采集方式，只要是大批量地采集互联网数据，获取到的往往是大量冗余且信息价值较低的数据。这不仅大量消耗数据源方的资源和服务性能，也大量消耗数据采集方的计算、存储、网络等资源。数据需求者往往是采取尽可能全量方式去采集数据并分析利用数据，使得数据分析者或数据利用方难以清晰地认知互联网大数据，即所需的数据在哪里、数据有多少、数据成分是什么等问题不清楚。本标准的数据资源分布图主要反映网络大数据的基本特征，为数据分析者提供网络大数据源的基本情况。

本标准提出原位虚拟数据中心体系结构及构建方法，旨在面向数据中心的数据采集与分析的需求，对互联网数据资源、互联网数据访问、互联网数据内容形成相对有序的有效的获取、分析和利用，从而为数据中心提供数据资源分布制导服务，为数据提供方提供数据访问协议。其特点包括以下几点：（1）原位虚拟数据中心相对于现有的数据中心来说是一种轻量型的数据中心，其自身并不采集和存储目标源全量数据，而是对互联网数据进行勘探，虚拟化互联网数据资源，并形成数据资源分布图。（2）原位虚拟数据中心的核心在于网络数据勘探器和数据资源分布图。通过网络数据勘探器对网站信息量和价值密度进行勘探，利用采用样本估计方法得到该网络站点的数据规模价值密度和分布情况等信息，并生成数据资源分布图。根据数据资源分布图，并通过数据资源获取制导服务向数据需求方法提供数据源及其互联网数据资源的大致分布情况和价值信息。

4 数据资源分布图结构



第0层节点（根节点）描述项：	数据分类方法	数据分类数量	访问限制	类别1指针	类别2指针	……	类别n指针	扩展项
第1层节点描述项：	数据模态数	限制命令	文本类指针	图像类指针	视频类指针	语音类指针	其他类指针	扩展项
第2层节点描述项：	数据站点数	限制命令	站点1指针	站点2指针	站点3指针	……	站点n指针	扩展项

第3层节点（数据节点）描述项：	数据位置	限制命令	数据量	数据成分	数据分布	数据时序性	访问命令参数	返回数据格式	扩展项

图 1 数据资源分布图结构

数据资源分布图的结构如图 1 所示，主要包括第 0 层节点（根节点）、第 1 层节点、第 2 层节点、第 3 层节点(数据节点)。其中，第 0 层节点（根节点）、第 1 层节点、第 2 层节点为初始化层节点，第 3 层节点(数据节点)为扩展层节点，4 层节点构成树形结构。具体结构及其构造描述如下：

(1) 第 0 层节点（根节点）主要包括：数据分类方法、数据分类数量、访问限制、类别 1 指针、类别 2 指针……、类别 n 指针、扩展项等描述。其中，数据分类方法项记录用于数据分类模型或方法；类别指针用于指向类别节点，即根节点的每个孩子节点为一个类别，扩展项用于信息扩充。

(2) 第 1 层节点主要包括：数据模态数、限制命令、文本类指针、图像类指针、视频类指针、语音类指针、其他类指针、扩展项等描述。数据模态数是指数据模态的分类数，一般情况指文本、图像、视频、语音以及其他等五种数据；文本类指针、图像类指针、视频类指针、语音类指针、其他类指针是记录指向孩子节点的链接指针，其孩子节点为某种数据模态的节点。

(3) 第 2 层节点主要包括：数据站点数、限制命令、站点 1 指针、站点 2 指针、……、站点 m 指针、扩展项等描述。数据站点数是指某类某种数据模态下的数据源站点的总个数，该数量同时表明其孩子的节点数；站点指针记录了其每个孩子节点。

第 3 层节点为数据节点， 主要包括：数据位置、限制命令、数据量、数据成分、数据分布、数据时序性、访问命令及参数、返回数据格式、扩展项等描述。数据位置记录了该数据源的站点位置；限制命令为访问该数据源的限制访问描述；数据量为该站点的数据数量，数据提供方提供（也可为空）；数据成分表明数据的组成元素；数据分布是数据的基本特征及其分布情况；数据时序性表明数据之间是否为时间序列关系；访问命令及参数记录访问该数据源的命令及其参数（也可为空）；返回数据格式是指所获取的数据的格式。

5 原位虚拟数据中心系统的框架

原位虚拟数据中心系统的架构，如图 2 所示。

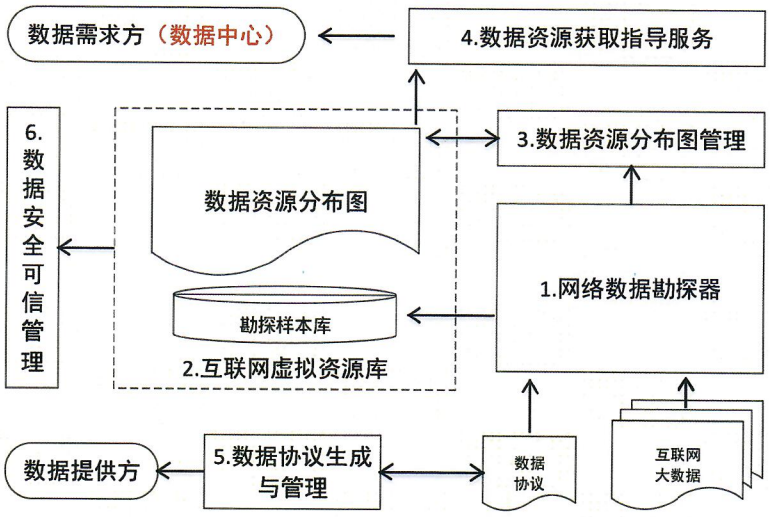


图 2 原位虚拟数据中心系统的框架

原位虚拟数据中心的体系结构，如图 2 所示，主要由网络数据勘探器、互联网虚拟资源库、数据资源分布图管理、数据资源获取制导服务、数据协议生成于管理、数据安全可信管理等子系统构成。具体子系统及构成原理如下：

(1) 网络数据勘探器。是原位虚拟数据中心的核心子系统之一，负责对互联网数据进行采样评估并生成数据资源分布图，具体包括：数据采样引导单元，用于根据所述数据提供方提供的数据访问协议文件，产生数据采样引导信息，以实现互联网 Web 数据采样引导和/或内部数据库应用程序编程接口采样引导；数据采样引导信息的数据结构表示为数据采样引导树和/或数据采样引导表；数据采样引导树是对互联网数据进行采样的引导信息；数

据采样引导表是通过应用程序编程接口访问网络站点的内部数据库的数据采样引导信息表；数据采样估算单元，用于根据数据采样引导树和/或数据采样引导表，采样抓取互联网数据至所述互联网虚拟资源库；同时进行互联网 Web 数据采样估算和/或内部数据库应用程序编程接口采样估算；属性信息包括数据类别、数据模态、数据量、数据成分、数据分布；数据资源分布图生成单元，用于根据互联网数据的属性信息以及数据采样引导树中访问限制，生成数据资源分布图。

(2) 互联网虚拟资源库。用于存储所述数据资源分布图及所述互联网数据勘探器采集的样本数据，主要包含数据资源分布图和勘探样本库。数据资源分布图是原位虚拟数据中心的核心数据结构组件，它反映了互联网数据的整体分布情况，包括数据位置、数据量、数据特征等信息，是大规模数据采集的指导信息表，如图 1 所示。勘探样本库是用于存储网络数据勘探器采集的小样本数据。

(3) 数据资源分布图管理子系统。是对数据资源分布图进行存储访问、更新等操作的管理系统。其中，所述数据资源分布图采用关系型或非关系数据库存储；数据资源分布图的访问按照树形结构进行访问。本标准中数据资源分布图管理的核心是数据资源分布图的动态更新方法，该方法将保证互联网虚拟资源库保持最新状态。

(4) 数据资源获取制导服务子系统。是根据资源分布图向数据需求方提供数据采集与挖掘的指导服务，以保证数据需求用户能高效地、有序地采集挖掘互联网数据及其进一步的分析。

(5) 数据协议生成与管理子系统。是根据互联网数据提供方所提供的数据访问协议、以及数据站点地图生成统一的数据访问协议文件，包括 Web 数据访问协议、互联网内部数据库访问协议等，并能够对这些协议提供管理功能，包括协议的发布、更新等。

(6) 数据安全可信管理子系统。用于对所述互联网虚拟资源库中虚拟数据资源进行数据安全管理，主要对虚拟数据资源的访问管理，包括数据隐私保护、数据访问权限等管理。

6 原位虚拟数据中心的核心构造流程

原位虚拟数据中心的核心构造主要包括网络数据勘探器、资源分布图构造及其管理。步骤如下：

步骤 1：根据数据协议和互联网大数据构建网络数据勘探器，详细过程如下：

a) 勘探器模块根据数据协议分 Web 页面数据和内部数据库 API 两种类型进行采样引导；

b) 分析数据采样引导的结果，针对 Web 页面数据建立引导树，对内部数据库建立引导表；

步骤 2：根据网络数据勘探器勘探的数据结果，构建互联网虚拟资源库，包含勘探样本库和数据资源分布图；

a) 勘探器模块根据数据采样引导的结果，对互联网大数据分 Web 页面数据和内部数据库两种类型进行数据采样估算，估算数据包含数据量，数据成分，数据分布情况等价值，构建勘探样本库；

b) 根据数据采样引导树/引导表和数据采样估算的结果生成数据资源分布图。

步骤 3：根据网络数据勘探的结果和数据资源分布图进行数据资源分布图的管理，详细过程如下：

a) 根据数据资源分布图，提供数据资源获取制导服务；

b) 根据数据协议，按不同数据提供方的不同生成并管理原位虚拟数据中心的数据协议；

c) 对互联网虚拟资源库，包含勘探样本库和数据资源分布图的维护和管理。

7 网络数据勘探器的构造及资源分布图的生成

网络数据勘探器主要由四部分组成：数据采样引导模块、数据采样引导树/引导表、数据采样估算模块、数据资源分布图生成模块。如图 3 所示，其工作原理是：

(1) 数据采样引导模块主要是根据数据提供方的相关数据访问限制，生成数据采样引导信息。主要分为两类引导：一类是 Web 页面数据采样引导，一类是内部数据库 API 采样引导。Web 页面数据采样引导主要是读取互联网中的数据爬取协议文件、站点地图文件，并按照一定的策略读取部分数据，生成数据采样引导树。数据采样引导树记录了可访问数据站点资源及其访问权限等。内部数据库 API 采样引导

主要是通过读取数据提供方提供的访问方式及访问限制的标准访问文件，生成数据采样引导树；若没有提供标准的访问限制文件，则人工配置标准访问文件，然后再生成数据采样引导树。

(2) 数据采样引导树/引导表是数据采样引导模块生成的数据采样引导信息数据结构。

a) 数据采样引导树是指对 Web 信息进行采样的引导信息，Web 页面数据采样引导模块生成，如图 4 所示。Web 数据采样引导树主要是一颗树形结构，根节点是网站的根目录节点，子节点是子网站的子目录节点，每个节点的描述项包括数据位置（数据所在的站点位置）、数据模态（文本、图像、视频、语音等）、数据勘探器名字、数据访问的限制命令、数据的时序特征、访问命令、命令参数、返回的数据格式（页面或 Jason 等数据格式）、扩展项（用于其他 Web 形式数据的扩展描述）。

b) 数据采样引导表是指对互联网上通过API接口访问内部数据库的数据采样引导信息表，表 1 所示。主要包括数据位置（数据所在的站点位置）、数据模态、数据勘探器名字、访问禁止/限制项、API 调用函数表（含参数、返回值）描述、数据的时序性、数据的分布性、数据是否在线/离线、扩展项。

(3) 数据采样估算模块根 Web 数据采样引导树和内部数据库 API 采样引导表，按照一定的策略（区间采样或点采样策略）抓取一定数量的数据存入互联网虚拟资源库；同时进行互联网 Web 数据采样估算和/或内部数据库应用程序编程接口采样估算，估算数据的类别、数据模态、数据量、数据成分、数据分布等。

(4) 数据资源分布图生成模块根据数据采样的分析结果，以及数据采样引导树中的访问限制，生成数据资源分布图。

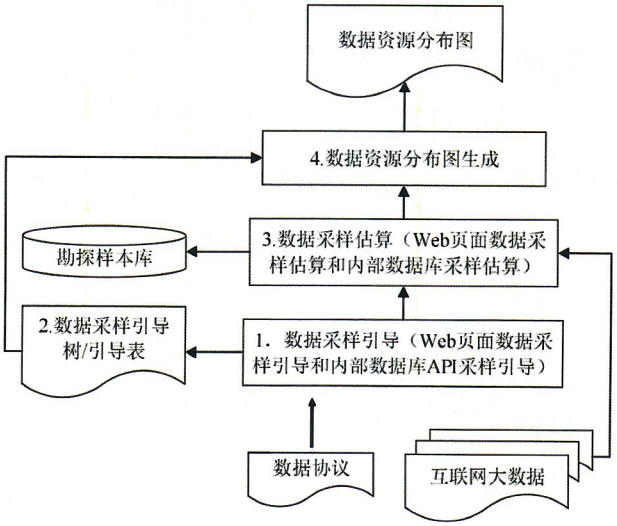


图 3 网络数据勘探器的构造

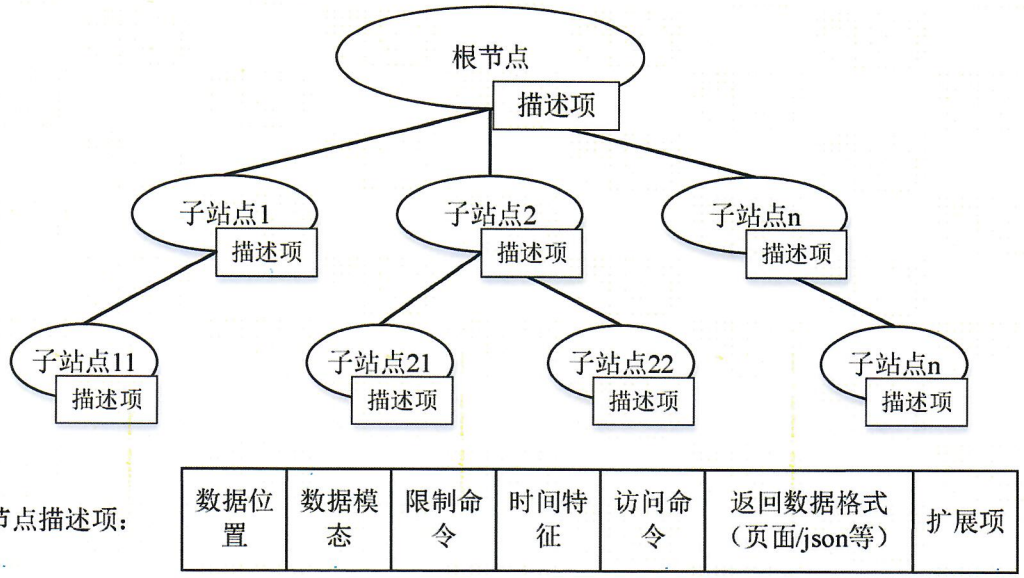


图 4 Web 数据采样引导树 Web-GuideTree 的构成模型

表 1 内部数据库 API 采样引导信息表 API-GuideList

资源位置	数据模态	数据勘探器名字	访问禁止/限制项	API 调用函数表（含参数、返回值）描述	数据的时序性	数据的分布性	数据是否在线/离线	扩展项

参 考 文 献

[1] GB/T 1.1-2009 标准化工作导则

[2] GB 4943.1-2011 信息技术设备 安全 第 1 部分：通用要求

[3] ISO/IEC 23006-4-2013 信息技术 多媒体服务平台技术 第 4 部分:基本服务

[4] ISO/IEC 23006-1-2013 信息技术 多媒体服务平台技术 第 1 部分:架构

[5] ISO/IEC 23006-5-2013 信息技术 多媒体服务平台技术 第 5 部分:服务聚合

[6] GB/T 29746-2013 实时交通信息服务数据结构

[7] JR/T 0011-2004 银行集中式数据中心规范

[8] GB 4943.23-2012/IEC 60950-23:2005 信息技术设备 安全 第 23 部分：大型数据存储设备

[9] ISO 5807-1985 信息处理 数据流程图、程序流程图、系统流程图、程序网络图和系统资源图的文件编制符号及约定

[10] GB/T 16680-1996 软件文档管理指南

[11] GB 33473-2016 即时通信业务 HI 接口总体技术要求

[12] GB/T 32399-2015 信息技术 云计算 参考框架

[13] GB/T 25000.51-2016 系统与软件工程 系统与软件质量要求和评价（SQuaRE）第 51 部分：就绪可用软件产品（RUSP）的质量要求和测试细则

[14] GB/T 32423-2015 系统与软件工程 验证与确认

[15] GB/T 25067-2016/ISO/IEC 27006:2011 信息技术 安全技术 信息安全管理体系审核和认证机构要求