GY

中华人民共和国广播电视和网络视听行业标准

GY/T 339. 1—2020

有线电视网络大数据技术规范 第 1 部分：通用要求

Technical specification for CATV’s big data—
Part 1: General requirements

2020－12－22 发布　　2020－12－22 实施

国家广播电视总局　　发 布

目　　次

前言 III
引言 IV
1　范围 1
2　规范性引用文件 1
3　术语和定义 1
4　缩略语 2
5　概述 2
6　数据采集接入要求 3
6.1　概述 3
6.2　基本要求 3
6.3　采集接入内容 3
6.4　数据表达 4
6.5　性能要求 4
6.6　交互与接口 4
7　数据处理要求 4
8　数据开放与服务要求 4
8.1　功能要求 4
8.2　开放内容 5
8.3　数据表达 5
8.4　性能要求 5
8.5　交互与数据接口 5
9　数据服务安全要求 5
9.1　概述 5
9.2　数据采集环节 5
9.3　数据传输环节 6
9.4　数据接入平台环节 6
9.5　数据存储环节 6
9.6　数据处理环节 6
9.7　数据开放环节 6
9.8　接口安全 6
9.9　个人信息隐私保护 6
附录A（规范性）　数据采集上报的交互过程与接口要求 8
A.1　交互过程 8

A.2 接口 9
A.3 OAuth2 密码模式认证过程 13
A.4 设备注册过程 15
A.5 控制信令消息格式 16
A.6 文件数据上报 18
附录 B（规范性） 大数据平台与应用系统的交互过程与接口要求 19
B.1 概述 19
B.2 认证接口 19
B.3 资源调用接口 20
参考文献 22

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件为GY/T 339《有线电视网络大数据技术规范》的第1部分。GY/T 339已经发布了以下部分：

——第 1 部分：通用要求；

——第 2 部分：平台要求；

——第 3 部分：数据规则。

请注意本文件的某些内容可能涉及专利。本文件发布机构不承担识别这些专利的责任。

本文件由全国广播电影电视标准化技术委员会（SAC/TC 239）归口。

本文件起草单位：国家广播电视总局广播电视规划院、中国电子技术标准化研究院、北京邮电大学、广州市诚毅科技软件开发有限公司、浩鲸云计算科技股份有限公司、北京东方国信科技股份有限公司、华数数字电视传媒集团有限公司、国家广播电视总局广播电视科学研究院、重庆有线电视网络股份有限公司、中国广播电视网络有限公司、北京歌华有线电视网络股份有限公司、广东省广播电视网络股份有限公司、湖北省广播电视信息网络股份有限公司、河北广电无线传媒有限公司、深圳市天威视讯股份有限公司、陕西广电网络传媒（集团）股份有限公司、陕西广信新媒体有限责任公司、贵州省广播电视信息网络股份有限公司、江苏省广电有线信息网络股份有限公司、北京海致星图科技有限公司、广西广电大数据科技有限公司、新疆广电网络股份有限公司。

本文件主要起草人：余英、韦安明、吴钟乐、张群、王洪波、刘智、王帅、刘敬玉、唐志燕、李庆国、聂明杰、邓向冬、曹志、王倩男、赵明、赵士原、欧阳峰、杨旭、沈文、唐永壮、董彬、刘军霞、胡其权、刘彦鹏、柳涛、杨晨、王飞、郑璐、林昕、梅杨、唐昊、陈昕、尹卓、曹燕明、诸葛海标、胡暐宸、张玮、刘晓敏、王欣然、曹阳、李海波、鞠宏、付晶、赵良福、苟明宇、杨敬一、王季友、刘艺兰、张城瑞、周传淯、傅力军、王瑶、范斐、孙嘉阳、张琦、陶宛昌、张君、王士刚、杨娟、李文、涂均、吕燕、刘波、彭宇涛、杨斌。

引　言

GY/T 339《有线电视网络大数据技术规范》规定了有线电视网络大数据技术规范的通用要求，包括大数据系统和数据服务的功能、性能、接口、安全等方面的要求，适用于有线电视网络大数据系统和业务的规划、设计、实施、验收、升级改造和运行维护。

GY/T 339共有三个部分。各部分简述如下。

——第1部分：通用要求。规定了有线电视网络大数据系统和数据服务的功能、性能、接口、安全等方面的要求。

——第2部分：平台要求。规定了有线电视网络大数据平台的结构和技术要求。

——第3部分：数据规则。规定了有线电视网络大数据的数据源、数据内容和数据表达规则。

有线电视网络大数据技术规范　第1部分：通用要求

1　范围

本文件规定了有线电视网络大数据技术规范的通用要求，包括大数据系统和数据服务的功能、性能、接口、安全等方面的要求。

本文件适用于有线电视网络大数据系统和业务的规划、设计、实施、验收、升级改造和运行维护。

2　规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；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 7408—2005　数据元和交换格式　信息交换　日期和时间表示法（ISO 8601:2000，IDT）

GB/T 35273—2017　信息安全技术　个人信息安全规范

GB/T 35295—2017　信息技术　大数据　术语

GY/T 339.2—2020　有线电视网络大数据技术规范　第2部分：平台要求

GY/T 339.3—2020　有线电视网络大数据技术规范　第3部分：数据规则

GD/J 074—2018　电视收视数据元素集规范

GD/J 075—2018　电视收视数据交换接口规范

IETF RFC 1952　GZIP文件格式规范4.3（GZIP file format specification version 4.3）

IETF RFC 3629　UTF-8，ISO 10646的一种转换格式（UTF-8，a transformation format of ISO 10646）

IETF RFC 8259　JSON数据交换格式（The JavaScript Object Notation(JSON) Data Interchange Format）

3　术语和定义

GB/T 35295—2017界定的以及下列术语和定义适用于本文件。

3.1

大数据　big data

具有体量巨大、来源多样、生成极快、多变等特征并且难以用传统数据体系结构有效处理的包含大量数据集的数据。

[来源：GB/T 35295—2017，定义2.1.1]

3.2

大数据参考体系结构　big data reference architecture

一种用作工具以便于对大数据内在的要求、设计结构和运行进行开放性探讨的高层概念模型。

[来源：GB/T 35295—2017，定义2.1.3]

3.3

数据中心 data center

由计算机场站（机房）、机房基础设施、信息系统硬件（物理和虚拟资源）、信息系统软件和信息资源（数据）等组成的实体。

3.4

大数据平台 big data platform

以大数据参考体系结构为功能基础的数据中心系统，在本文件中，指集成了大数据采集接入、存储、处理、分析、共享，以及各类配套功能组件及基础设施的数据处理系统。

3.5

大数据系统 big data system

以大数据参考体系结构为基础的数据处理系统，在本文件中，指由大数据平台、数据源、数据采集终端、网关以及相关辅助功能组件构成的数据处理系统。

[来源：GB/T 35295—2017，定义2.1.14]

3.6

数据采集终端 data collection terminal

一种部署在数据源实现数据规范收集汇总的软件组件或实体设备。

3.7

OAuth2 The OAuth 2.0 Authorization Framework

通过该框架，允许第三方应用程序通过IETF RFC 6749规定的方法获取HTTP服务或行为的有限制的访问权限。

注：OAuth2是由IETF RFC 6749描述的“OAuth 2.0授权框架”的简称。

4 缩略语

下列缩略语适用于本文件。

API 应用程序编程接口（Application Programming Interface）

BSS 业务支撑系统（Business Support System）

HTTP 超文本传输协议（HyperText Transfer Protocol）

HTTPS 安全超文本传输协议（Secure Hypertext Transfer Protocol）

JSON JavaScript对象标记（JavaScript Object Notation）

MSS 管理支撑系统（Management Support System）

OSS 运营支撑系统（Operation Support System）

SQL 结构化查询语言（Structured Query Language）

URI 统一资源标识符（Uniform Resource Identifier）

UTF-8 8位通用字符集转换格式（8-bit Unicode Transformation Format）

5 概述

本文件将大数据系统的结构用图1的形式进行建模，系统中包括数据采集接入、数据处理、数据应用等部分。本文件为大数据系统的通用要求部分，规定数据源、数据采集接入、数据处理、数据开放服务、数据服务安全及个人信息隐私保护的基本要求，以及数据采集终端与平台、平台与应用系统之间的交互和接口。

数据采集终端均通过网关与平台进行数据交换。对于不同的应用场景，网关可表现为不同的形式，一种为集成在数据采集终端内部的组件，一种为独立于数据采集终端和大数据平台的实体设备，也可能是集成在大数据平台内部的一个组件。为了便于定义大数据业务流程，本文件不区分网关的具体实现形式和位置，仅将其当成一个实现格式规范化转换和执行通信功能的功能实体。

大数据系统模型各单元之间有依赖关系，例如数据采集终端执行数据采集流程并上传数据到网关，网关接收数据采集终端的数据并封装成规范格式后上传到大数据平台，大数据平台接受终端和网关的认证注册，向终端和网关下发操作指令，同时为各类大数据应用提供开放的数据服务。

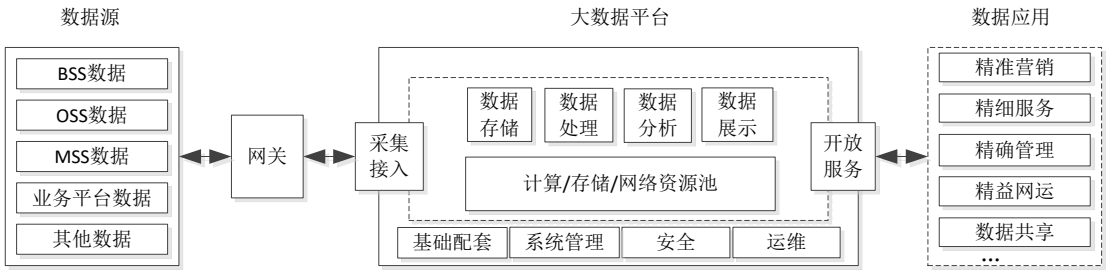


图1　大数据系统模型

6　数据采集接入要求

6.1　概述

数据采集接入指大数据平台将外部数据纳入大数据平台内部的过程。

6.2　基本要求

要求如下：

a）　应具备定时采集接入数据的功能；
b）　应具备定量采集接入数据的功能；
c）　宜提供图形化的数据采集接入配置或管理界面；
d）　应具备采集接入结构化数据的功能；
e）　应具备采集接入半结构化数据的功能；
f）　应具备采集接入非结构化数据的功能；
g）　应具备采集接入实时在线数据的功能；
h）　应具备采集接入离线数据的功能；
i）　应具备主动采集接入和被动接收数据的功能；
j）　宜采取措施维持数据源与接入服务器时钟的同步。

6.3　采集接入内容

要求如下：

a）　应具备采集接入 GY/T 339.3—2020 规定的各类基础数据的功能；

b） 应具备采集接入 GD/J 074—2018 规定的收视评价基础数据的功能。

6.4 数据表达

要求如下：

a） 应支持采集接入按 GY/T 339.3—2020 规定格式表达的数据；
b） 应支持采集接入按 GD/J 075—2018 规定的格式表达的符合 GD/J 074—2018 规定的数据。

6.5 性能要求

要求如下：

a） 采集接入实时数据时，应具备实时传输、接入全网实时数据的能力，实时数据应在不超过 3s 内从数据源到达大数据平台的数据存储系统；
b） 采集接入 BSS 离线数据时，应具备 1h 内完成当日增量数据的传输、接入的能力；
c） 采集接入 OSS 离线数据时，应具备 1h 内完成当日增量数据的传输、接入的能力；
d） 采集接入 MSS 离线数据时，应具备 1h 内完成当日增量数据的传输、接入的能力；
e） 采集接入其他离线数据时，应具备 1h 内完成该类数据当日增量的传输、接入的能力。

6.6 交互与接口

要求如下：

a） 数据源或数据采集终端与大数据平台间的交互应通过网关进行；
b） 数据源数据采集终端与大数据平台间的交互宜符合附录 A 的要求。

7 数据处理要求

大数据平台接入数据后，应具备数据处理功能，要求如下：

a） 应具备数据预处理功能，可对存放在文件系统和数据库中的数据进行抽取、清洗、转换等操作后加载到数据分析等处理模块；
b） 应具备存储处理后的数据的功能，可为其他处理模块提供上传、下载、查看、删除、权限管理等操作，并通过冗余备份等机制提供安全的数据存储；
c） 应具备计算资源配置、调度和回收管理功能，支持水平扩展计算框架、调度任务和管理任务优先级；
d） 应具备数据分析功能，支持非结构化数据、实时数据、结构化数据的分析，可提供 SQL 分析、跨数据源关联分析、机器学习等数据分析功能，并提供对大数据平台外开放分析服务的 API；
e） 应具备接入、处理符合 GY/T 339.3—2020 规定的数据的能力；
f） 数据处理环节的功能、性能、数据访问、系统和数据管理、基础配套、安全可靠性和运行维护等应满足 GY/T 339.2—2020 的要求。

8 数据开放与服务要求

8.1 功能要求

要求如下：

a） 应具备对外提供数据服务的功能，如提供数据交换、共享、发布等功能；
b） 宜具备对外开放数据存储、分析、展示功能；

c） 应具备管理数据服务用户注册、授权、计费功能，可监看和审计用户的行为；
d） 应提供开放的数据服务接口，如提供用户注册 API、服务应用 API，宜提供二次开发接口，允许用户自定义业务；
e） 应能提供至少 1 年内的 BSS、OSS、MSS 数据；
f） 其他功能应满足 GY/T 339.2—2020 的要求。

8.2 开放内容

要求如下：
a） 宜具备开放数据终端接入的原始数据的功能；
b） 应具备开放 GY/T 339.3—2020 定义的数据集的功能；
c） 应具备开放处理或分析数据结果的功能；
d） 宜具备开放数据存储、处理和分析能力的功能。

8.3 数据表达

开放的数据集，收视数据应按GD/J 075—2018规定的格式表达，其他数据格式应符合GY/T 339.3—2020的规定。

8.4 性能要求

大数据平台对外提供数据开放服务时，性能要求如下：
a） 应用系统向大数据平台发送的注册或认证消息，大数据平台的平均响应时间应小于 1s；
b） 应用系统向大数据平台请求数据交换的响应时间、交换的数据包大小、共享的数据容量、多数据源访问响应时间、并发用户数等性能应满足 GY/T 339.2—2020 的要求。

8.5 交互与数据接口

平台与应用系统之间的交互宜符合附录B的要求。

9 数据服务安全要求

9.1 概述

在开展数据服务的过程中，应采取措施确保大数据系统的数据采集、传输、存储、处理、数据开放的安全。

9.2 数据采集环节

要求如下：
a） 应确保数据采集的合法性和正当性；
b） 应按照 GY/T 339.3—2020 规定的范围采集数据，按照 GY/T 339.3—2020 的要求规范数据格式，明确采集频度；
c） 应通过对组件、终端、设备等采集终端进行必要的技术控制，如在部署前对采集终端进行采集能力认证、计量等，确保数据的完整性、一致性和真实性；
d） 应明确数据收集和获取过程中个人信息和重要数据的知悉范围和安全管控措施，确保采集数据的合法性、完整性和真实性；
e） 应采取脱敏、加密等技术或管理措施确保采集过程中涉及的个人信息和重要数据不被泄露；

f） 应对数据采集行为进行权限管理。

9.3 数据传输环节

要求如下：

a） 应对敏感数据进行加密传输；

b） 宜对重要数据进行加密传输。

9.4 数据接入平台环节

要求如下：

a） 应建立数据采集终端/网关接入平台的身份识别与鉴别策略、权限分配策略和相关操作规程，建立访问控制时效管理机制；

b） 应具备数据接入的安全审计功能。

9.5 数据存储环节

要求如下：

a） 应采取冗余存储或多副本存储措施，确保数据的可用性；

b） 应采取校验等检验和容错技术措施，确保多副本数据存储的一致性；

c） 宜具备数据归档离线存储功能；

d） 应确保存储技术架构具备加密敏感数据、重要数据的能力。

9.6 数据处理环节

数据处理发生在大数据平台内部，要求如下：

a） 应确保处理过程中发生数据迁移、变换时的真实性、完整性和一致性；

b） 应确保数据处理过程操作行为可审计，数据质量可溯源，重要数据受保护。

9.7 数据开放环节

要求如下：

a） 应建立应用系统接入平台的身份识别与鉴别策略、权限分配策略和相关操作规程，建立访问控制时效管理机制；

b） 应建立应用系统访问平台数据、服务的控制策略；

c） 应具备数据交换、共享、发布的安全审计功能。

9.8 接口安全

要求如下：

a） 应制定数据服务接口安全控制策略，明确规定使用服务接口的安全措施，如身份鉴别、授权策略、访问控制机制、签名、时间戳、安全协议等；

b） 应制定数据服务接口安全规范，包括接口名称、接口参数、对接安全要求等，具备限制或过滤不安全参数的能力；

c） 应具备数据服务接口访问审计功能；

d） 宜在数据采集终端与平台、平台与应用系统间采用安全通道或加密通道进行接口调用。

9.9 有线电视网络用户个人信息隐私保护

在采集、传输、存储、展示、共享、管理有线电视网络中涉及个人的信息数据时，在以下方面应遵循GB/T 35273—2017的要求：

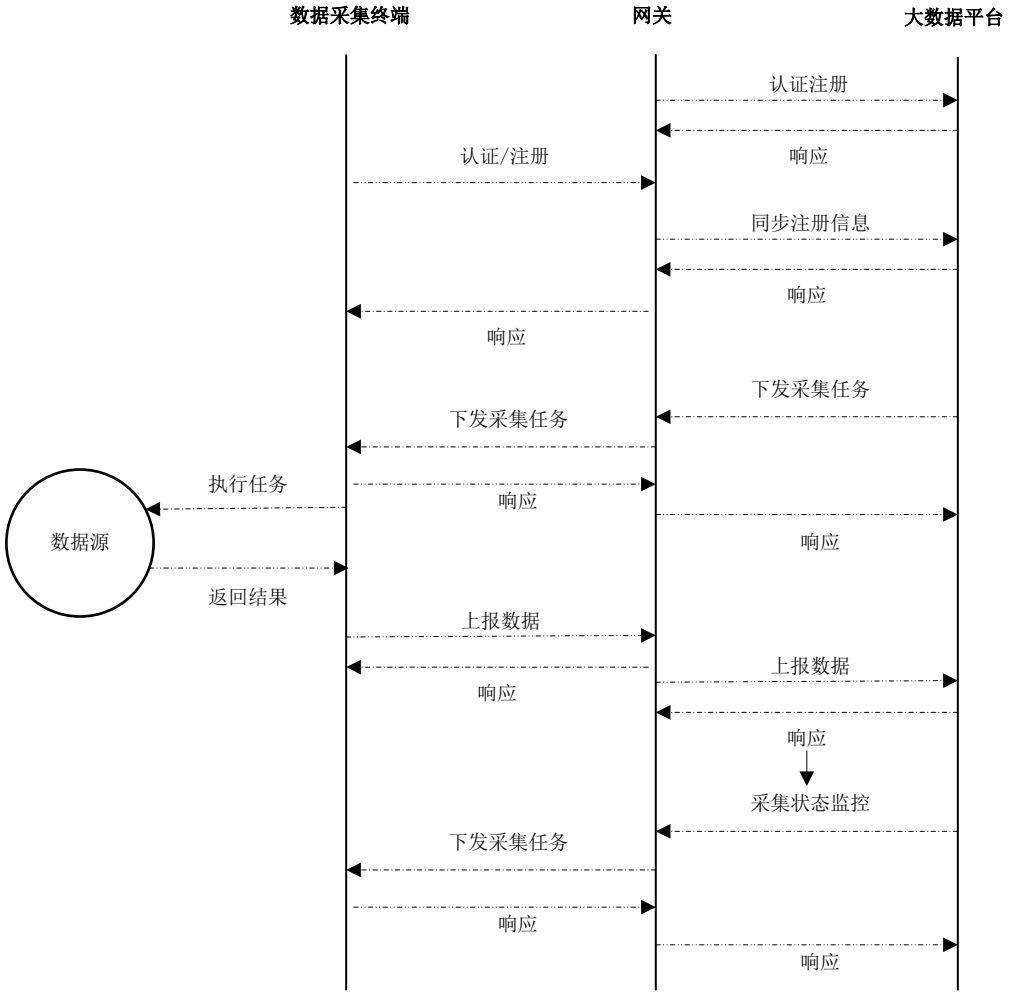
a) 个人信息的采集；
b) 个人信息的保存；
c) 个人信息的使用；
d) 个人信息的委托处理、共享、转让和公开披露；
e) 个人信息安全事件的处置；
f) 从事个人信息处理、保管、使用等工作的组织和个人的管理。

附 录 A
（规范性）
数据采集上报的交互过程与接口要求

A.1 交互过程

A.1.1 概述

为了实现大数据系统的整体功能，大数据平台、数据采集终端、网关单元之间应按照本文件约定的机制相互协调工作，图A.1展示了单元间的必要协调机制及通信过程，包括认证注册、控制信令下发和数据上报。



图A.1 大数据系统各单元间的通信过程

A.1.2 交互过程

各单元间协调工作时，信令交互过程和要求如下：

a) 网关应可通过规范接口，向大数据平台提交设备认证及注册请求，成功后方可继续后续操作，否则无法成为大数据系统中的可管理的一个单元；
b) 数据采集终端应可通过规范接口，向网关提交设备认证及注册请求，成功后方可继续后续操作；
c) 网关在收到数据采集终端的注册请求时，应可将注册信息同步到大数据平台，同步成功后数据采集终端成为大数据系统中的可管理的一个单元；
d) 大数据平台应可对数据采集终端进行统一管理，包括向终端下发采集任务、查看终端采集状态等信令，这些信令应由相应的通信单元转发送达数据采集终端；
e) 数据采集终端收到信令后，应按要求执行相应的操作，例如收到数据采集信令，则开始执行采集操作，并将采集到的数据上报给网关；
f) 网关应能接收并处理数据采集终端上报的消息和数据，对于格式不符合规范要求的，网关进行格式转换、聚合和封装处理后上传到大数据平台；
g) 大数据平台应可接收并处理网关上报的数据。

A.2 接口

A.2.1 概述

考虑到各数据采集终端所使用的采集技术的差异，本文件仅规定数据交换接口的共性要求。

基于大数据的数据量大，部分数据存在实时接收处理需求，本章定义的接口消息格式和传输协议兼顾考虑传输效率和实时性的要求，此外还考虑：

a) 独立性：不以具体的操作系统或程序语言限制采集数据的格式、实现方式；
b) 标准性：保持采集数据交换格式和数据封装方式的一致性，确保基本数据项的完整；
c) 开放性：采用基础通信协议和开放、通用的数据标识方式，避免数据交换过程产生歧义；
d) 兼容性：兼容多种数据来源格式；
e) 可扩展性：在终端、平台、机构等发生变化时仍可进行接口适配。

A.2.2 认证注册

A.2.2.1 概述

按照本文件的描述，系统中存在两类设备认证注册接口，分别适用于两类场景：数据采集终端向网关认证注册，网关向大数据平台认证注册。系统设计实现时，应根据设备角色选择相应的接口。

A.2.2.2 数据采集终端向网关进行设备认证注册

数据采集终端进行设备认证注册时，要求如下：

a) 数据采集终端应先向网关完成设备认证，才能进行后续交互；
b) 设备认证所使用的凭据，应是能够标识设备身份的信息，例如产品序列号、硬件序列号等，相关认证凭据，应预先在认证服务端录入，以便认证时进行比对；
c) 认证方式宜采用 OAuth2 认证框架；
d) 认证模式宜采用 OAuth2 协议中的密码模式，认证过程应符合 A.3 的规定；
e) 认证通过后，数据采集终端应立刻向网关发起设备注册请求，注册过程见 A.4；
f) 网关收到数据采集终端的注册请求并处理完成后，应将同步注册信息实时到大数据平台，同步过程见 A.4。

A.2.2.3 网关向大数据平台进行设备认证注册

网关进行设备认证注册时，要求如下：

a) 网关应先向大数据平台完成设备认证，才能进行后续交互；
b) 设备认证所使用的凭据，应是能够标识设备身份的信息，例如产品序列号、硬件序列号等，相关认证凭据，应预先在认证服务端录入，以便认证时进行比对；
c) 认证方式宜采用 OAuth2 认证框架；
d) 认证模式宜采用 OAuth2 协议中的密码模式，认证过程见 A.3；
e) 认证通过后，网关应向大数据平台实时发起设备注册请求，注册过程见 A.4。

A.2.3 控制信令

为保证控制信令在各单元间传递，数据采集终端与网关之间应始终维持一条控制信令传输通道，每个网关和大数据平台之间同样应维持一条控制信令传输通道。

控制信令的传递方式如下，数据采集终端可以通过控制信令通道上报消息到网关，然后经网关中转，通过网关和大数据平台之间的控制信令通道，最终到达大数据平台。同样，大数据平台主动发送的消息也应经网关中转后到达数据采集终端。

控制信令通道应通过长连接来实现，如基于TCP的全双工通讯协议WebSocket，该协议支持持久连接，能在数据采集终端和大数据平台之间保持长连接，并且连接双方都可以作为消息发送方主动发起消息。WebSocket协议净荷为消息，格式描述应符合IETF RFC 8259规定的JSON字符串的要求，字符编码应符合IETF RFC 3629定义的UTF-8编码。

控制消息应具有统一的头部信息，各字段定义应符合表A.1的规定。

表 A.1 控制消息头部信息

参数名称	字段	类型	参数说明
提供方名称	ProviderName	String	自定义
提供方代码	ProviderID	String	标识提供数据方设备的唯一代码
消息类型	MsgType	Int	传输消息的类型
消息动作	MsgAction	Int	传输消息对应的动作
上报时间	MsgTime	String	数据上报的时间戳，精确到秒，符合GB/T 7408—2005的5.4.1中“日期和日的时间的组合”的“完全表示法”，数据格式为yyyymmddhhmmss

基本的控制信令消息应包括心跳状态（消息类型1000）、采集任务下发（消息类型2000）、数据采集终端状态查询（消息类型2001）、采集任务状态查询（消息类型2002），格式见A.5，也可自定义控制信令，但应遵循本条所规定的技术要求。

A.2.4 数据上报

A.2.4.1 概述

数据采集终端执行采集任务后，获取到的结果通过数据上报接口发送到大数据平台。

数据采集终端生成的数据宜分为数据采集终端到网关、网关到大数据平台两步上报。

A.2.4.2 数据采集终端到网关的数据上报

数据采集终端应通过网关中转上报数据，不应直接将数据直接上报到大数据平台。

为了尽少占用网关资源，数据采集终端到网关的数据上报宜采取短连接，数据发送完毕后立刻关闭网络连接，节省连接资源。

数据传输协议应采用HTTP/HTTPS协议，URI中的域名为网关地址，采用HTTP POST报文，消息格式应符合IETF RFC 8259规定的JSON字符串规范，字符编码应符合IETF RFC 3629定义的UTF-8编码。

数据上报消息应附加表A.2定义的通用字段，用于数据描述。

表 A.2　终端到网关的通用字段定义

参数名称	字段	类型	参数说明
提供方名称	ProviderName	String	自定义
提供方代码	ProviderID	String	标识提供数据方设备的唯一代码
数据类型	DataType	String	传输数据的类型，如OA数据、收视数据等
上报时间	PostTime	String	数据上报的时间戳，精确到秒，符合GB/T 7408—2005“基本格式”中的“完全表示”方法，数据格式为yyyymmddhhmmss

以下为数据上报消息的参考格式，以直播收视行为数据为例：

```
POST /data HTTP/1.1
Host: data.server.com
Authorization: Bearer 2YotnFZFEjr1zCsicMWpAA
Content-Type: application/json

{
    "DeviceID": "123456X",
    "DeviceRegionID": 123456,
    "Time": "20190323160015",
    "ServiceType": 2,
    "ActionType": 1,
    "object": {
        "ChannelID": "3201001001",
        "ChannelName": "某广播电视台综合频道",
        "Status": 1,
        "EnterType": 2,
        "PlayStatus": 1
    }
}
```

A.2.4.3　网关到大数据平台的数据上报

网关应对收到的数据采集终端数据进行聚合和格式转换等处理后，再发送到大数据平台。

为了尽少占用网关资源，网关到大数据平台的数据上报宜采取短连接，数据发送完毕后立刻关闭网络连接，节省连接资源。

数据传输协议应采用HTTP/HTTPS协议，URI中的域名为大数据平台地址，采用HTTP POST报文，消息格式应符合IETF RFC 8259规定的JSON字符串规范，字符编码应符合IETF RFC 3629定义的UTF-8编码。

网关作为HTTP请求的发起方，应对POST请求的正文内容（JSON字符串）进行数据压缩处理，数据压缩格式应采用符合IETF RFC 1952定义的GZIP压缩格式，同时应在请求头部中添加字段“Content-Encoding: gzip”。

大数据平台作为HTTP请求的接收方，应支持对GZIP格式数据的解压缩。

数据上报消息应附加表A.3定义的通用字段，用于数据描述。

表 A.3　网关到大数据平台的通用字段定义

参数名称	字段	类型	参数说明
网关名称	GatewayName	String	自定义
网关代码	GatewayID	String	标识网关的唯一代码
聚合时间	AggrTime	String	数据聚合的时间戳，精确到秒，符合GB/T 7408—2005“基本格式”中的“完全表示”方法，数据格式为yyyymmddhhmmss

以下为数据上报消息的参考格式，为多条数据以数组的形式聚合到一条数据消息中：

```
POST /data HTTP/1.1
Host: data.server.com
Authorization: Bearer 2YotnFZFEjr1zCsicMWpAA
Content-Type: application/json

{
    "GatewayName": "XXXX网关",
    "GatewayID": "1111",
    "AggrTime": "20190323160025",
    "object": [
        {
            "ProviderName": "某公司",
            "ProviderID": "123456X",
            "DataType": "Live",
            "PostTime": "20190323160015",
            "object": {
                "ChannelID": "3201001001",
                "ChannelName": "某广播电视台综合频道",
                "VideoFormat": 2,
                "ProgramID": "1301019901",
                "ProgramName": "某节目",
                "ProgramLength": "005700",
                "ProgramPlaydate": "20180104",
                "ProgramPlayTime": "172500"
              }
        },
        {
            "ProviderName": "XXXX公司",
            "ProviderID": "123456X",
            "DataType": "VOD",
            "PostTime": "20190323160015",
```

```
                "object": {
                "ChannelID": "3201001002",
                "ChannelName": "某广播电视台综合频道",
                       "VideoFormat": 2,
                       "ProgramID": "1301019901",
                       "ProgramName": "某节目",
                       "ProgramLength": "005700",
                       "ProgramPlaydate": "20180104",
                       "ProgramPlayTime": "172500"
                    }
            },
            {
                "ProviderName": "XXXX公司",
                "ProviderID": "123456X",
                "DataType": "VOD",
                "PostTime": "20190323160015",
                "object": {
                    "ChannelID": "3201001002",
                    "ChannelName": "某广播电视台综合频道",
                    "VideoFormat": 2,
                    "ProgramID": "1301019901",
                    "ProgramName": "某节目",
                    "ProgramLength": "005700",
                    "ProgramPlaydate": "20180104",
                    "ProgramPlayTime": "172500"
                  }
            }
        ]
}
```


根据数据源的实时性要求不同，数据上报方式可选用实时交换模式、准实时交换模式或非实时交换模式，不同的模式要求如下：

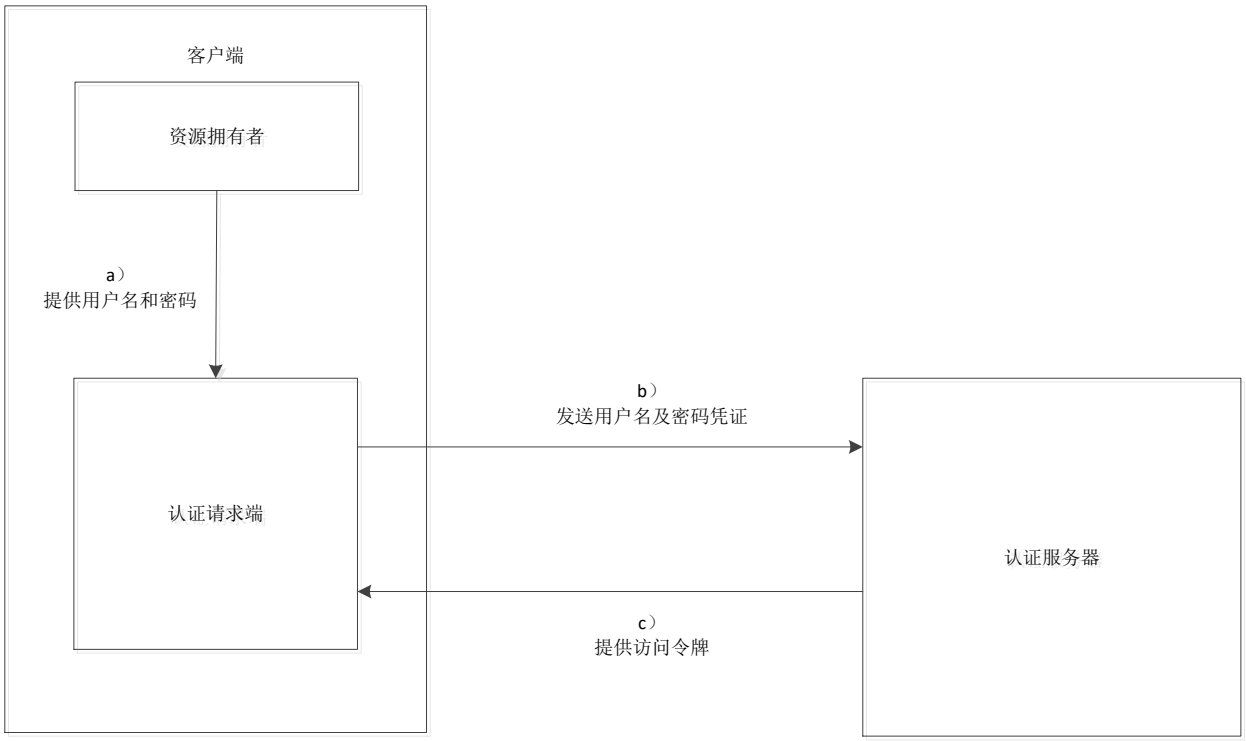
a) 实时交换模式下，数据交换双方应采用 HTTP/HTTPS 传输协议进行数据交换；
b) 准实时交换模式下，数据提供方应与数据接收方进行协商，双方在规定的时间将一段时间内生成的采集数据通过 HTTP/HTTPS 进行数据交换；
c) 非实时交换模式下，数据提供方宜以文件的方式提交到数据接收方，相关要求见 A.6，或者通过存储介质传递等其他方式进行数据交换。

A.3 OAuth2 密码模式认证过程

认证过程应符合图A.2的规定，客户端采用密码模式，通过以下过程完成认证：

a) 客户端（同时也是资源拥有者）使用标识自身唯一身份的凭据（如序列号）作为用户名，同时将该用户名进行 MD5 摘要提取操作（32 位大写格式）后得到的字串，作为该用户名对应的密码；

b) 客户端将用户名和密码发给认证服务器，向后者请求令牌；
认证服务器确认无误后，向客户端提供访问令牌。



图A.2 基于密码模式的 OAuth2 认证流程

认证请求采用URI路径和加密的HTTPS协议进行传输。URI中的域名为提供认证服务的服务端地址，如“www.server.com”。

HTTP采用POST报文，报文体中的消息以x-www-form-urlencoded编码方式进行编码。

HTTP请求需包含如下参数：

——grant_type：表示授权类型，此处的值固定为"password"，必选项；

——username：表示用户名，必选项；

——password：表示用户的密码，必选项；

——scope：表示权限范围，可选项。

一个客户端设备认证报文示例如下：

```
POST /auth HTTP/1.1
Host: www.server.com
Authorization: Basic czZCaGRSa3FOMzpnWDFmQmFOM2JW
Content-Type: application/x-www-form-urlencoded
grant_type=password&username=121000004000012553X&password=656EB371DA1C51D2E5241C03A617BC3C
```

认证服务提供方需对请求方的认证请求进行处理鉴权，并给出相应的响应信息；请求方根据请求返回的响应结果，判定是否认证成功，认证成功才能继续后续流程，否则终止操作。

认证服务返回的结果状态码含义，见IETF RFC 2616。

一个认证成功返回结果的报文示例如下：

```
HTTP/1.1 200 OK
Content-Type: application/json;charset=UTF-8
Cache-Control: no-store
Pragma: no-cache
{
    "access_token": "2YotnFZFEjr1zCsicMWpAA",
    "token_type": "example",
    "expires_in": 3600,
    "refresh_token": "tGzv3JOkF0XG5Qx2TlKWIA",
    "example_parameter": "example_value"
}
```

A.4 设备注册过程

设备注册请求采用URI路径和加密的HTTPS协议进行传输。URI中的域名为提供注册服务的服务端地址，如“www.server.com”。

HTTP采用POST报文，报文体中的消息格式符合IETF RFC 8259规定的JSON字符串的要求，字符编码应符合IETF RFC 3629定义的UTF-8编码。

HTTP请求报文体中需包含如下字段：

——ProviderID：表示请求方设备ID，必选项；

——ProviderType：表示请求方角色类型，“terminal”表示注册角色为数据采集终端，“gateway”表示注册角色为网关，必选项。

其余字段为可选项，可根据具体需求定制。

一个客户端设备注册报文的示例如下：

```
POST /register HTTP/1.1
Host: www.server.com
Authorization: Bearer 0b79bab50daca910b000d4f1a2b675d604257e42
Content-Type: application/json
{
    "ProviderID": "12100000400012553X",
    "ProviderType": "terminal"
}
```

注册服务提供方需对请求方的注册请求进行处理，并给出相应的响应信息；请求方根据请求返回的响应结果，判定是否注册成功，注册成功才能继续后续流程，否则终止操作。

注册服务返回的结果状态码含义，见IETF RFC 2616。

一个设备注册成功返回结果的报文示例如下：

```
HTTP/1.1 200 OK
Content-Type: application/json;charset=UTF-8
Cache-Control: no-store
Pragma: no-cache
{
    "status": 200,
```

```
        "id": 1000
    }
```

A.5 控制信令消息格式

A.5.1 概述

本章为心跳状态、采集任务下发、数据采集终端状态查询、采集任务状态查询控制信息的格式定义。

A.5.2 心跳状态

数据采集终端和大数据平台之间通过心跳信息来保持存活性和状态更新，数据采集终端需每隔一定时间，向大数据平台主动上报自己的心跳信息，消息类型代码为1000。

一条心跳信息的参考格式如下：

```
    {
        "ProviderName": "某公司",
        "ProviderID": "123456X",
        "MsgType": 1000,
        "MsgAction": 0,
        "MsgTime": "20190323160015",
        "object": {
            "HeartBeatTime": "20190323160015"
        }
    }
```

A.5.3 采集任务下发

大数据平台通过向数据采集终端发送类型为2000的消息，来实现对终端下发采集任务的操作，参考格式如下：

```
    {
        "ProviderName": "大数据平台",
        "ProviderID": "1",
        "MsgType": 2000,
        "MsgAction": 0,
        "MsgTime": "20190323160015",
        "object": {
            "TaskID": 12345,
            "...": "..."
        }
}
```

采集服务器收到下发任务消息后，执行任务下发相关逻辑，并将执行结果反馈到大数据平台，回应消息参考格式如下：

```
    {
        "ProviderName": "某公司",
```

```
    "ProviderID": "123456X",
    "MsgType": 100,
    "MsgAction": 0,
    "MsgTime": "20190323160015",
    "object": {
        "Status": 200  //参考IETF RFC 2616状态码
    }
}
```

A.5.4 数据采集终端状态查询

大数据平台通过向数据采集终端发送类型为2001的消息来获取终端的状态，参考格式如下：

```
{
    "ProviderName": "大数据平台",
    "ProviderID" : "1",
    "MsgType": 2001,
    "MsgAction": 0,
    "MsgTime": "20190323160015"
}
```

终端收到该类型的消息后，需对自身的系统状态进行获取，然后回应给大数据平台，参考格式如下：

```
{
    "ProviderName": "某公司",
    "ProviderID": "123456X",
    "MsgType": 110,
    "MsgAction": 0,
    "MsgTime": "20190323160015",
    "object": {
        "Status": 200,
        "CPU": 0.1,
        "Memory": 0.6
    }
}
```

A.5.5 采集任务状态查询

大数据平台通过向数据采集终端发送类型为2002的消息来查询终端执行任务的状态，参考格式如下：

```
{
    "ProviderName": "大数据平台",
    "ProviderID": "1",
    "MsgType": 2002,
    "MsgAction": 0,
```

```
        "MsgTime": "20190323160015"
    }
```

终端收到该类型的消息后，需对自身的采集任务状态进行获取，然后回应给大数据平台，参考格式如下：

```
{
    "ProviderName": "某公司",
    "ProviderID": "123456X",
    "MsgType": 111,
    "MsgAction": 0,
    "MsgTime": "20190323160015",
    "object": {
        "Status": 200,
        "Tasks": [
            {
                "TaskID": 123456,
                "Lasting": 300, //秒
                "DataSize": 1024 //字节
            },
            {
                "TaskID": 234567,
                "Lasting": 300, //秒
                "DataSize": 1024 //字节
            }
        ]
    }
}
```

A.6 文件数据上报

在一些非实时等应用场合，为了提高数据上报效率，数据采集终端先将采集到的数据保存到文件中，满足上报条件时再集中上传。

以文件形式进行数据上报的基本规则如下：

a) 数据记录应以文本文件格式保存，文件中的内容应为数据记录的简单堆叠，以便于数据解析、拆分、合并及进行分布式处理，也利于 Loader 工具将其加载到数据库中；

b) 如果文本文件中包含了若干条没有上下文关联的记录，则宜将同一数据类型，例如同一地区、同一类设备、同一组用户、同一时间段等数据有规律地组织到同一个文件中，以提高数据处理效率；

c) 文件宜根据数据来源、地理位置、生成日期等规律按目录分级放置或上传；

d) 文件名宜包含数据提供者、数据生成日期、数据所属地区、用户分组、终端类型、终端分组等信息。

附 录 B
（规范性）
大数据平台与应用系统的交互过程与接口要求

B.1 概述

在大数据平台与应用系统的交互过程中，定义应用系统为资源调用者，而大数据平台为资源提供者，下文据此定义进行描述。

资源调用者和资源提供者的交互过程中主要包含两类开发接口：认证接口，以及资源调用接口。

在资源调用者向资源提供者调用上述的开发接口之前，资源提供者需提供基于WEB的应用注册页面，以便资源调用者注册自己的应用，并且为应用申请唯一的标识：client_id和client_secret。

一个典型的应用系统与大数据平台的交互流程见图B.1。

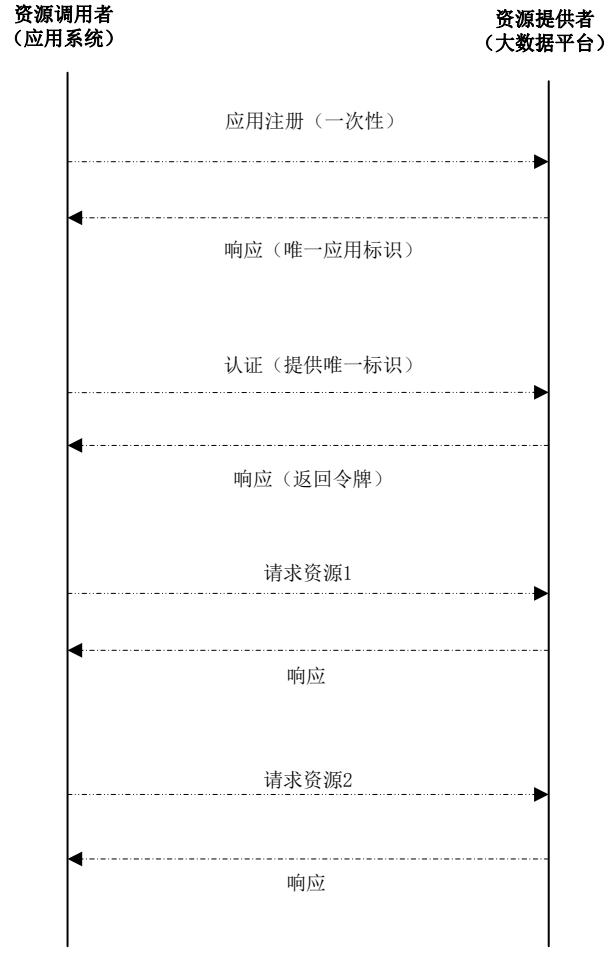


图 B.1 资源调用者和资源提供者交互流程

B.2 认证接口

资源调用者在调用资源提供者的资源和能力集之前，需先通过资源提供者方面的认证，确认当前接入连接为合法的应用接入。

认证基于OAuth2认证机制，采取Client Credentials方式授权方案，其认证流程可简要描述如下：待认证的应用客户端使用自己的client证书(如client_id及client_secret组成的http basic验证码）来获取access token。

认证请求采用URI路径和加密的HTTPS协议进行传输。URI中的域名为资源提供者中提供认证服务的服务端地址，如“api.server.com”。

HTTP采用POST报文，报文体中的消息以x-www-form-urlencoded方式进行编码。

HTTP请求需包含如下参数：

——grant_type：表示授权类型，此处的值固定为“client_credentials”，为必选项；

——client_id：表示应用的唯一ID，为应用注册成功后上文中获取和使用的应用的ID；

——client_secret：表示应用ID对应的密钥，为应用注册后上文中获取的应用的secret；

——scope：表示权限范围，为可选项。

下面是一个资源调用者的认证报文示例：

```
POST /api/auth HTTP/1.1
Host: api.server.com
Authorization: Basic czZCaGRSa3FOMzpnWDFmQmFOM2JW
Content-Type: application/x-www-form-urlencoded
grant_type=client_credentials&client_id=MzRNTxgske3QRf5Yj69&client_secret=
a3OCAbcbDuuGLdHLeyRaZk1tq5
```

资源提供方需对资源调用方的认证请求进行处理鉴权，并给出相应的响应信息；资源调用方根据认证请求返回的响应结果，判定是否认证成功，认证成功才能继续后续资源调用流程，否则终止操作。

认证服务返回的结果状态码含义，见IETF RFC 2616。

一个认证成功返回结果的报文示例如下：

```
HTTP/1.1 200 OK
Content-Type: application/json;charset=UTF-8
Cache-Control: no-store
Pragma: no-cache
{
    "access_token": "2YotnFZFEjr1zCsicMWpAA",
    "token_type": "example",
    "expires_in": 3600,
    "refresh_token": "tGzv3JOkF0XG5Qx2TlKWIA",
    "example_parameter": "example_value"
}
```

B.3 资源调用接口

资源调用方需要调用资源提供方的相关资源或能力集时，需通过相应的资源调用接口完成。

资源调用接口应当是一组接口，符合特定格式，并且通过不同的URI字段来区分资源调用的类型。

本章仅描述资源调用接口应当遵循的格式规范，不定义某个具体的资源调用接口。

资源调用接口采用URI路径和加密的HTTPS协议进行传输。URI中的域名为提供注册服务的服务端地址，如“api.server.com”。

请求资源的类型，需通过URI中的具体路径来区分，如获取BSS数据的URI：/resource/bss。

HTTP采用POST报文，报文体中的消息格式符合IETF RFC 8259规定的JSON字符串的要求，字符编码应符合IETF RFC 3629定义的UTF-8编码。

HTTP请求报文体中需包含如下字段：

Authorization：认证请求中获得的access_token，格式为Bearer access_token。

其余字段为可选项，可根据具体请求的资源类型进行添加。

请求资源的其他选项，需在请求body中进行定制，如定制请求资源的起始时间信息，限制请求资源的大小等。

一个资源调用请求报文的示例如下：

```
POST /api/resource/bss HTTP/1.1
Host: api.server.com
Authorization: Bearer 2YotnFZFEjr1zCsicMWpAA
Content-Type: application/json
{
    "StartTime": "20181001080000",
    "StopTime": "20181031080000",
    ……
}
```

资源提供方需对资源调用方的资源调用请求进行处理，并给出相应的响应信息；资源调用方根据请求返回的响应结果，判定是否请求成功，如请求成功，则对所请求的资源进行接收处理，否则终止本次请求操作。

资源调用服务返回的结果状态码含义，见IETF RFC 2616。

一个资源调用请求成功返回结果的报文示例如下：

```
HTTP/1.1 200 OK
Content-Type: application/json;charset=UTF-8
Cache-Control: no-store
Pragma: no-cache
{
    "status": 200,
    "type": "bss",
    "data": {
     ……
     ……
    }
}
```

参 考 文 献

[1] GB/T 35274—2017 信息安全技术 大数据服务安全能力要求
[2] IETF RFC 2616 HTTP/1.1（Hypertext Transfer Protocol—HTTP/1.1）
[3] IETF RFC 6749 The OAuth 2.0 Authorization Framework