GY

中华人民共和国广播电视和网络视听行业标准

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

Technical specification for CATV’s big data—
Part 2: Platform's 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

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2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、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、北京邮电大学、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、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、王洪波、王飞、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、周传淸、傅力军、王瑶、范斐、孙嘉阳、张琦、陶宛昌、张君、王士刚、杨娟、郑璐、林昕、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2部分：平台要求

1　范围

本文件规定了有线电视网络大数据平台的结构和技术要求，还规定了对有线电视网络大数据采集、接入、存储、处理、分析、展示和开放服务的技术要求。

本文件适用于有线电视网络数据的采集、接入、存储、处理、分析、展示和开放服务，还适用于指导有线电视网络运营机构开展大数据平台的规划设计、实施、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 22239—2019　信息安全技术　网络安全等级保护基本要求

GB/T 35295—2017　信息技术　大数据　术语

GB/T 35589—2017　信息技术　大数据　技术参考模型

GB 50174—2017　数据中心设计规范

GY/T 337—2020　广播电视网络安全等级保护定级指南

GD/J 075—2018　电视收视数据交换接口规范

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

数据采集终端 data collection terminal

一种部署在数据生成节点的、实现数据规范收集汇总和处理的软件组件或实体设备。

3.6

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，以及相关辅助等功能组件构成的数据处理系统。

[来源：GB/T 35295—2017，定义2.1.14]。

3.7

元数据 meta data

一种带有数据类型、编码、名称、业务描述等属性的，可用于描述数据产品特征的数据单元。

3.8

框架 framework

一种由数据的采集、接入、存储、处理、分析、服务等功能组件构成的集合。

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）
CEP 复杂事件处理（Complex Event Processing）
CPU 中央处理器（Central Processing Unit）
GPU 图形处理器（Graphic Processing Unit）
HDD 硬盘驱动器（Hard Disk Drive）
IO 输入输出（Input and Output）
MPI 消息传递接口（Message-Passing-Interface）
MSS 管理支撑系统（Management Support System）
NoSQL 非关系型的数据库（Not only SQL）
OLAP 在线分析处理（On-Line Analysis Processing）
OSS 运营支撑系统（Operation Support System）
RAID 独立磁盘冗余阵列（Redundant Arrays of Independent Drives）
SQL 结构化查询语言（Structured Query Language）

SSD 固态盘（Solid State Disk）

XML 可扩展标记语言（Extensible Markup Language）

5 概述

大数据平台实现对运营机构经营和系统运行维护过程中产生的各类大数据的采集、接入、处理、存储、分析、展示、共享和管理，为大数据消费者提供数据和服务，以及为运营机构间的数据交换提供统一接口。图1采用GB/T 35589—2017中关于大数据参考体系架构的定义，采用角色、活动、组件等逻辑构件描述有线电视网络大数据平台（以下简称大数据平台或平台）的组成和业务逻辑。关于大数据参考体系架构，以及角色、活动、组件的描述见附录A。

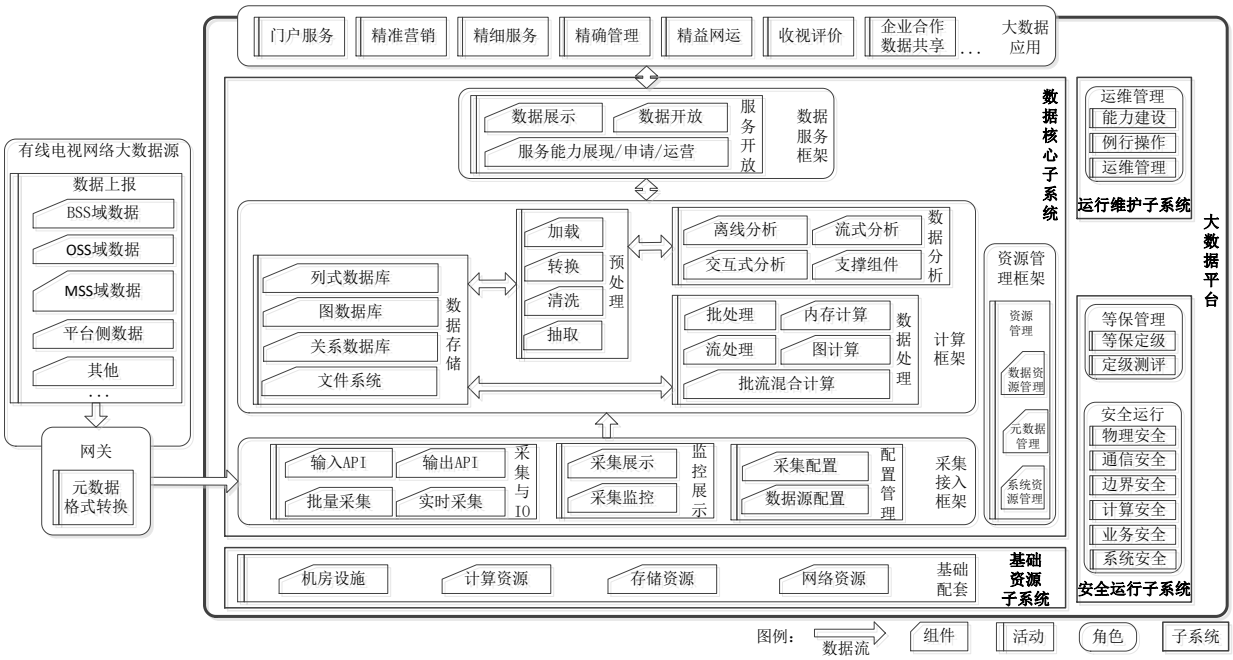


图1 大数据平台示意图

大数据平台由数据核心子系统、基础资源子系统、安全运行子系统和运行维护子系统组成。数据核心子系统是大数据平台的核心组成部分，由数据采集接入、数据存储、数据处理、数据分析，数据服务、资源管理等框架组成，负责实现数据输入、计算处理和输出功能，以及对外开放大数据平台的数据服务功能。基础资源子系统为大数据平台提供机房、计算资源、存储资源、网络资源等基础服务，安全运行子系统和运行维护子系统为数据处理提供基本保障。

大数据平台应支持采集、接入、处理运营机构的BSS、OSS、MSS数据，广告、媒资、用户收视行为、用户体验数据，以及网管、门户网站等数据，具体支持的数据内容如图2所示。上传到大数据平台的数据，其格式和交互方式应与GY/T XXX.1—XXXX的要求相符，例如大数据平台应具备采集接入符合GD/J 075—2018要求的数据。

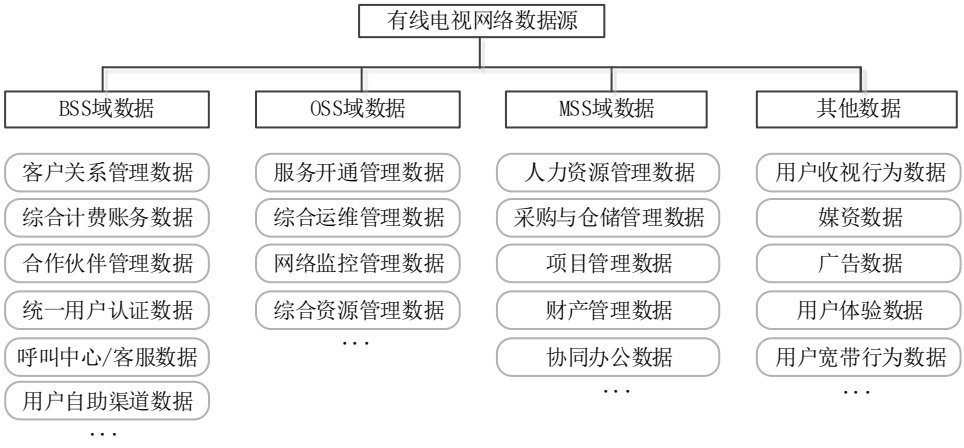


图2　数据内容

6　数据核心子系统技术要求

6.1　数据采集接入

6.1.1　基本要求

要求如下：

a）　应支持结构化数据、半结构化数据、非结构化数据的批量导入；
b）　宜支持从文件服务器的多个目录并发导入数据；
c）　应支持离线数据导入功能，延时应不超过 24h；
d）　应支持实时采集；
e）　应支持定时采集，如可根据预设的频率（小时、天等），自动对选定的数据源进行增量或者全量采集或导入；
f）　应支持对采集对象进行远程配置，如远程设置采集对象的采集频度、采集时间、采集数据量等；
g）　应支持采集对象的注册、发布、使用授权、变更、注销等管理；
h）　应支持手动控制采集任务的执行过程；
i）　宜支持图形化的数据导入配置界面。

6.1.2　输入输出 API

要求如下：

a）　应为数据采集终端或网关提供数据输入 API；
b）　应为计算框架提供数据输出 API。

6.1.3　采集接入监控

要求如下：

a）　应支持监控采集对象的运行状态；
b）　应支持监控采集网络的运行状态，如采集对象所在网络的通断；
c）　应支持监控采集任务的执行情况，如支持了解采集任务执行时间、任务进度、已采集数据量等；
d）　应支持监控采集服务器 CPU、内存、硬盘和网络适配器的使用情况，如果采集服务器部署在虚拟机中，应支持监控虚拟机的 CPU、内存、硬盘和网络适配器的使用情况；

e） 应支持监控采集任务队列状态；
f） 应支持异常情况预警，包括采集任务执行失败、采集节点状态异常、网络异常等；
g） 应支持通过任务状态分析、数据完整性分析等方法监控非联网设备的数据采集情况；
h） 应支持可定制的监控告警，如提供电子邮件、短信、智能终端 APP 等告警方式；
i） 监控应不影响采集任务的正常执行；
j） 宜图形化展示监控数据。

6.1.4 性能要求

要求如下：
a） 采集接入实时数据时，应具备实时接入全网实时数据的能力，并在不超过 3s 内将接收的实时数据存入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接入的能力；
e） 采集接入其他离线数据时，应具备 1h 内完成当日增量数据接入的能力。

6.2 数据存储

6.2.1 基本要求

要求如下：
a） 应支持存储结构化数据、半结构化和非结构化数据；
b） 应支持数据上传、数据下载、目录查看、目录创建、目录删除、权限修改等操作；
c） 应具备标准、开放的数据访问 API，以支持对数据的操作；
d） 应支持对用户访问进行授权；
e） 应具备数据加载工具或功能，使系统和关系型数据库、其他文件系统之间可进行数据和文件交换；
f） 应支持存储调度，可按用户计划对存储节点的迁移、扩容、复制、更改、删除等进行自动操作；
g） 应支持文件分级存储，如单机级、跨服务器级、跨机柜级、跨数据系统级；
h） 应支持为用户设置不同的数据存放策略；
i） 应支持为单个用户提供逻辑存储空间；
j） 应在多用户之间设置数据隔离机制；
k） 应支持在数据源端去重处理；
l） 应支持表管理功能；
m） 应支持负载均衡，负载均衡切换过程中，业务应不中断；
n） 应支持对关系型数据库的不同数据实例制定独立的数据备份策略；
o） 宜支持数据自动备份和手动备份；
p） 宜支持数据批量更新、删除等数据管理操作；
q） 宜支持流式实时数据入库和实时查询。

6.2.2 文件系统

要求如下：
a） 应支持文件的上传、下载、读写、复制、移动、删除、访问控制等；
b） 应具备文件容错机制和系统高可用机制，包括数据块备份、系统快速恢复等功能；

c） 应支持文件数据的校验和同步，保证数据的完整性与一致性；
d） 应支持分布式文件系统的弹性扩展，支持节点动态添加和删除；
e） 应支持压缩和加密存储的数据；
f） 应支持快速检索，如数据资源的统一检索、编目、增加和删除；
g） 应支持文件的搜索、批量操作、回收站、快照；
h） 宜支持小文件打包为大文件集中存储；
i） 宜支持存储配额管理，可基于目录存储空间及文件数量进行配额控制；
j） 宜支持分级存储的功能，如根据数据的使用热度、时延要求等特性将数据分别存储在SSD、HDD等介质中。

6.2.3 数据库支持

6.2.3.1 关系数据库

要求如下：
a） 应支持结构化数据存储机制，实现数据存储的可扩展性；
b） 应支持多表关联；
c） 应支持数据存储一致性检查，实现数据的完整性与一致性；
d） 宜支持行列混合存储，支持表按行或列格式组织存储；
e） 宜支持行列转换。

6.2.3.2 列式数据库

要求如下：
a） 应支持以键值形式进行数据存储；
b） 应支持基于表、列族和列的用户权限管理，权限管理操作包括读、写、创建等；
c） 应支持对数据库中的列进行加密；
d） 应支持数据的备份与恢复，包括库级别的备份和恢复，并提供备份恢复进展、历史记录查看等功能；
e） 宜支持多级索引；
f） 宜支持将多个具有类似功能或存在关联关系的业务表进行合并存储。

6.2.3.3 图数据库

要求如下：
a） 应支持由节点及边组成的数据模型；
b） 应支持图查询、图遍历、图分析、图挖掘等；
c） 应支持主流开发接口；
d） 应支持单节点、多节点多层关系扩线查询，支持广度优先、深度优先、最短路径、最优路径遍历搜索算法；
e） 宜支持顶点、属性的继承操作；
g） 宜支持长任务异步会话机制。

6.2.4 性能要求

要求如下：
a） 列式数据库存储，实时处理响应时间应小于1s，一亿条记录的批处理响应时间应小于10s；

b） 缓存数据库，响应时间应小于 10ms，IO 能力应不小于 10000 条每秒；
c） 负载均衡切换时间应小于 10s；
d） 存储处理设备 CPU 忙时平均利用率宜小于 70%；
e） 存储处理设备内存忙时平均利用率宜小于 80%；
f） 底层存储的冗余保护能力应不低于 RAID“0+1”方式；
g） 恢复备份系统中的数据库时，恢复时长应小于 6h；
h） 备份数据库到备份系统时，每天的备份时间应小于 6h。

6.3 数据处理

6.3.1 基本要求

要求如下：
a） 分布式数据库应支持 CPU、内存、GPU 等异构资源调度和配置；
b） 分布式数据库应支持计算框架的水平扩展；
c） 应支持任务优先级调度，能定义不同优先级的任务，使得后启动的高优先级任务能够获取运行中的低优先级任务释放的资源；
d） 应支持对全局资源的集中管理；
e） 应支持静态资源分配策略和动态资源分配策略；
f） 分布式数据库应提供与组织相匹配的层次结构，应支持多层次的队列资源管理，队列的资源应严格隔离，队列获得的资源应不超过分配给该队列的上限；
g） 分布式数据库应支持弹性资源与抢占，即当有空闲资源时，租户可以使用超过其配置资源，以提高系统整体的吞吐量；当系统繁忙，其他租户无法拿到配置应得的资源时，当前租户超过配置部分的资源可以被其他租户抢占，以保证各租户的服务质量；
h） 分布式数据库应支持资源管理、作业调度和数据加载，以及各种分布式计算框架的调度；
i） 分布式数据库宜支持按照任务间的依赖关系自动调度任务，以提高处理系统的自动化程度；
j） 分布式数据库宜支持根据作业需求动态分配计算资源，自动管理回收资源；
k） 分布式数据库宜支持自动完成作业调度，并支持作业内多任务以无回路有向图形式描述的依赖关系；
l） 分布式数据库宜支持复杂任务的调度，如支持深度学习的训练、MPI 任务。

6.3.2 批处理

要求如下：
a） 应支持多种数据类型的离线分析，包括结构化、半结构化、非结构化数据；
b） 应支持离线计算任务进度与状态的实时上报；
c） 应支持多节点离线任务联动执行；
d） 应支持多种语言分析任务的开发接口；
e） 应支持作业调度；
f） 应支持分散-聚集的处理方式；
g） 应支持作为一个批处理计算框架运行在分布式资源管理之上。

6.3.3 流处理

要求如下：

a) 应支持从数据源中获得实时消息数据，完成高吞吐、低延迟的实时计算，并将结果输出到消息队列或者进行持久化；
b) 应支持对消息处理任务进行创建、浏览、中止、激活、去激活等操作，并记录用户级别的操作到审计日志中；
c) 应支持创建滑动窗口方式的实时分析任务，其时间窗口大小应可调；
d) 应支持通过 SQL 或者类 SQL 接口对数据进行操作；
e) 应支持容错性，在出现故障时具备容错处理机制。

6.3.4 图计算

要求如下：

a) 应内置图数据查询类 API；
b) 应支持以同步计算模型或异步计算模型的迭代算法；
c) 应支持明细数据全量导入、增量导入以及自定义导入；
d) 应支持内存计算和索引，支持在线图分析和查询；
e) 应支持基于属性图模型的图数据表达，包含结点/边上的标签和属性类型定义；
f) 应支持内置常用图指标计算功能，以描述图的拓扑结构特征；
g) 应支持实现水平扩展的分布式图计算和查询；
h) 应支持图数据的并发查询。

6.3.5 内存计算

要求如下：

a) 应支持基于内存的分布式计算；
b) 应支持水平扩展；
c) 应支持自动负载均衡；
d) 应支持多种数据类型的离线分析，包括结构化数据、半结构化数据、非结构化数据的离线分析；
e) 宜支持高度抽象算子，以快速构建分布式数据处理应用；
f) 宜支持标准 SQL 语法；
g) 宜支持与非关系型数据库对接，以在不迁移数据的前提下读取非关系型数据库中的数据并进行计算。

6.3.6 批流融合计算

要求如下：

a) 应支持批流融合统一查询 SQL 语言；
b) 应支持多种场景下的流式 SQL，如位置信息分析等；
c) 应支持机器学习、图计算；
d) 应支持时间窗口算法，包括跳跃窗口、滑动窗口等；
e) 宜支持基于 SQL 语言的批、流数据模式识别；
f) 宜提供批流融合计算引擎；
g) 宜支持事件驱动的流处理，以降低处理延迟；
h) 宜支持处理乱序事件流、窗口计算、CEP 等；
i) 宜支持迭代处理。

6.3.7 性能要求

要求如下：

a） 批处理集群容量的水平扩展能力应不小于 1000 个节点；

b） 在正常情况下，数据处理系统文件系统的 I0 请求响应时间应小于 500ms；

c） 在正常情况下，数据处理系统处理消息的响应延迟应小于 500ms。

6.4 数据分析

6.4.1 概述

数据分析过程由数据预处理、支撑能力、分析与挖掘过程、流程编排环节组成。通过数据分析过程，将数据信息转换为知识。数据分析的基本过程包括：对原始数据进行预处理，加载数据到分析系统，通过预置的分析工具库，执行检索查询、机器学习、统计分析、可视化等操作。一个完整的数据分析系统，通常还包括配置管理和任务流程编排功能。

6.4.2 数据预处理

6.4.2.1 数据抽取

要求如下：

a） 应支持按照需求抽取存放在文件系统、数据库中的文件或数据；

b） 对结构化、半结构化、非结构化数据，应支持不同的抽取方法；

c） 应支持全量抽取及增量抽取模式；

d） 应支持主动抽取和被动追加方法；

e） 应支持定时批量抽取；

f） 宜支持分布式数据抽取，数据抽取过程支持负载均衡。

6.4.2.2 数据清洗

要求如下：

a） 应支持检查数据一致性，支持清洗掉不一致的数据；

b） 应支持处理无效值，包括对无效值的删除、修正等；

c） 应支持处理缺失值，包括对缺失值的填充或缺失值对应数据条目的删除等；

d） 应支持处理重复值，包括对重复数据值的合并或删除等；

e） 应支持对比清洗前后的数据，方便使用者检验清洗效果；

f） 宜支持逻辑矛盾和关联性验证，支持清洗不合理的数据。

6.4.2.3 数据转换

要求如下：

a） 应支持对清洗后的数据按照分析模块的要求进行转换操作，支持结构化数据的列转换、行转换和表转换；

b） 宜支持非结构化数据、半结构化数据的结构化处理；

c） 宜支持对文本、网页类数据的规范化处理，将文档类数据转化成单一规范形式；

d） 宜支持对语音/音频数据的识别处理，将语音的词汇内容转换为计算机可读的输入；

e） 宜支持对图片中的内容转换为字符文本，支持提取图像信息。

6.4.2.4 数据加载

要求如下：

a) 应支持把经过清洗和转换之后的数据加载到分析系统，为分析功能模块提供数据。
b) 宜支持全量加载，即按照加载目标结构，将转换过的数据输入到目标结构中。若目标结构中已存在数据，则装入新数据进行覆盖。
c) 宜支持增量加载，即如果目标结构中已经存在数据，在保存已有数据的基础上增加新的数据。当输入的数据记录与已经存在的记录重复时，则丢弃新输入的数据，或将输入的记录作为副本进行增加。
d) 应支持实时加载或者批量加载。

6.4.3 分析支撑

6.4.3.1 检索查询

6.4.3.1.1 查询接口

要求如下：
a) 应支持标准的数据库查询接口；
b) 应支持 RESTful API 查询接口。

6.4.3.1.2 查询优化

要求如下：
a) 应支持基于规则的查询优化；
b) 应支持建立数据索引，提高查询效率；
c) 应支持数据分片和多副本技术，优化查询速度；
d) 应支持通过 SQL 进行复杂条件高并发查询；
e) 应支持精确查询和模糊查询；
f) 宜支持二级索引。

6.4.3.2 机器学习

6.4.3.2.1 数据管理

要求如下：
a) 应支持将输入数据划分为训练集、验证集和测试集；
b) 应支持导入和导出机器学习模型，支持导入训练和验证过的模型到分析系统中，以及导出训练所得的模型；
c) 宜支持多种数据模型的融合应用。

6.4.3.2.2 算法

要求如下：
a) 应支持回归与分类算法；
b) 应支持聚类算法；
c) 应支持协同过滤算法；
d) 应支持降维算法；
e) 应支持频繁模式挖掘算法；
f) 宜具备机器学习流程的其他组件，如特征提取、特征转换、特征选择、模型选择、交叉验证、模型调优等；

g） 宜支持通过二次开发增加新的指令算子。

6.4.3.2.3 任务管理

要求如下：

a） 应支持对不同的机器学习算法编排不同的数据分析流程，以得到适用于特定分析场景的机器学习模型；

b） 宜支持对机器学习任务进行分布式计算。

6.4.3.2.4 模型评估

宜提供用于评估算法模型的模块。

6.4.3.3 统计分析

要求如下：

a） 应支持基本的数值分析统计，如最大值、最小值、求和、总数等统计量；

b） 应支持数据集中趋势的分析统计，如平均数、中位数、众数等统计量；

c） 应支持数据离散程度的分析统计，如极差、方差、标准差等统计量；

d） 应支持分析多个随机变量的关系，如协方差、相关系数等统计量；

e） 宜支持自定义统计分析模板，并可保存常用的分析方案为模板。

6.4.3.4 可视化

要求如下：

a） 应支持将常见的数据源的格式作为输入；

b） 应支持可视化展示高维数据；

c） 具备可视化工具库，要求如下：

1） 应支持柱状图；
2） 应支持饼图；
3） 应支持折线图；
4） 应支持表格；
5） 宜支持散点图；
6） 宜支持雷达图；
7） 宜支持网络图；
8） 宜支持时间线；
9） 宜支持热力图；
10） 宜支持地图；
11） 宜支持桑基图；
12） 宜支持双轴图；
13） 宜支持箱线图；
14） 宜支持与算法模型评估相关的可视化工具库。

6.4.4 数据分析

6.4.4.1 分析模式

6.4.4.1.1 离线数据分析

要求如下：

a） 应支持结构化查询语言；
b） 应支持对离线数据的分布式分析；
c） 应具备支持第三方应用的标准接口；
d） 应支持分布式计算或并行计算等计算框架；
e） 应支持对海量工作任务的切分和分布式调度；
f） 应支持集成第三方机器学习算法库；
g） 宜支持使用内存或 SSD 存储作为缓存；
h） 宜支持对文本类、音视频类以及图像类数据的分析；
i） 宜支持对关系型数据库和大数据存储系统中的数据源进行交叉查询、聚合、关联操作；
j） 宜支持使用 GPU 对特定算法进行加速。

6.4.4.1.2 流数据分析

要求如下：

a） 应支持按时间切片后进行批量处理；
b） 应支持基于事件触发的流式处理；
c） 应支持关于实时流的数据统计；
d） 应支持流式数据的排序；
e） 应支持数据流与静态表之间的关联；
f） 应支持多个数据流的关联处理；
g） 采用滑动窗口方式的实时分析任务，其时间窗口大小应可调；
h） 宜支持实时数据的分组；
i） 宜支持分析任务优先级调度；
j） 宜支持对文本类、音视频类以及图像类数据的分析。

6.4.4.1.3 交互式联机分析

要求如下：

a） 应支持通过 SQL 或类 SQL 语言，对数据进行分布式的联机分析，如 OLAP 等；
b） 应支持通过结构化查询语言对数据进行即席（ad-hoc）查询；
c） 应支持利用可视化中间件对数据分析结果进行展示；
d） 应支持在交互式分析过程中定义计算公式和参数配置；
e） 应支持交互式分析过程的自动保存和回退等操作；
f） 应支持在交互式分析过程中对分析结果的保存和发布；
g） 应支持基于 OLAP 的交互式数据分析；
h） 宜支持对文本类、音视频类以及图像类数据的分析。

6.4.4.2 分析类型

6.4.4.2.1 预测型分析

要求如下：

a） 应支持趋势预测、回归分析等分析方法；
b） 应将准确率数值化；
c） 宜通过可视化的方式展示分析结果；

d） 应支持存储和发布训练好的模型。

6.4.4.2.2 描述型分析

要求如下：

a） 应支持相关关系分析方法；
b） 应支持可视化展示样本数据的分析结果，支持展示模型的训练效果，支持存储和发布训练好的模型；
c） 宜优化分析结果的呈现，提高用户体验。

6.4.5 流程编排

要求如下：

a） 应支持持久化保存流程编排结果；
b） 应支持跟踪计算或任务的执行状态，可给出异常任务告警；
c） 应支持工作流的调度触发机制，可配置触发时间或触发事件，可配置调度的启动时间和执行周期；
d） 应支持输出任务执行状态到日志；
e） 宜支持流程编排操作界面可视化，宜通过拖拉方式编排和修订流程；
f） 宜提供操作工作流的启动和停止的界面；
g） 宜支持并行执行多流程任务；
h） 宜支持通过数据管道实现任务串联；
i） 宜支持多人协同操作。

6.4.6 性能要求

要求如下：

a） 应支持万亿级数据联表，每天 IO 能力达到 PB 级；
b） 分析系统的数据吞吐能力应不小于 400MB/s。

6.5 数据展示

6.5.1 功能要求

要求如下：

a） 应具备数据展示模板，提供模板继承和整合功能；
b） 应提供数据展示界面和数据展示服务接口；
c） 数据展示服务应能兼容不同数据格式；
d） 宜通过缓冲、内存计算、压缩传输等方法，提高展示的响应速度；
e） 应支持结构数据（包括多维数据）、半结构数据、非结构数据的展示；
f） 应具备可扩展性，可通过二次开发，支持新的数据类型和可视化技术；
g） 应支持以下展示形式：
    1） 结构化数据下支持几何图展示，如仪表盘、饼状图、柱状图、曲线图、曲面图、雷达图等；
    2） 支持专业报表、即席报表、企业级复杂报表、自定义报表等报表展示形式；
    3） 支持假设分析和多维分析等数据挖掘的展示；
    4） 支持多种可视化图表的展示。

6.5.2 性能要求

要求如下：

a) 支持数据联表数量不少于 100,000,000 条；
b) 在一亿条数据记录规模下，SQL 查询平均响应时间应小于 5s；
c) 在一亿条数据记录规模下，NoSQL 的平均响应时间应小于 1s。

6.6 数据开放

6.6.1 功能要求

要求如下：

a) 应具备对外提供数据服务的功能，如向用户提供数据服务产品、处理用户对数据服务的申请、进行用户授权管理，以及服务计费、监控和审计等；
b) 应提供开放的数据访问 API；
c) 应提供数据分发功能；
d) 应支持按模板打包分发；
e) 应提供按需的数据存取访问服务；
f) 应允许用户配置和管理数据共享服务，如数据提取服务、数据发送服务等；
g) 宜提供数据分发二次开发接口，允许用户基于开发接口自定义业务；
h) 应支持对数据开放服务的管理和监控，如可管理数据服务的用户权限，查看运行日志，统计服务性能等。

6.6.2 性能要求

要求如下：

a) 在批量实时数据交换场景下，集群数据吞吐不低于 200MB/s 或 20 万条数据记录每秒时，单条数据记录平均响应时间不大于 100ms；
b) 应支持开放不小于 10TB 的数据容量；
c) 在提取多种数据源时，响应时间（用户向数据源发出请求到开始获得数据时间）应小于 30s；
d) 系统数据吞吐能力应不小于 400MB/s；
e) 支持的并发用户数应不小于 1000 个。

6.7 资源管理

6.7.1 数据资源管理

6.7.1.1 数据保护策略

要求如下：

a) 应支持数据分类、分级管理，可针对不同类别和级别的数据采取不同的保护措施；
b) 应支持数据安全标记，可按安全标记进行授权和访问控制；
c) 应在数据采集、存储、处理、分析等环节支持数据分类和分级，并应确保各环节对不同类别和级别的数据采取的保护策略是一致的；
d) 应在数据清洗和转换过程中对重要数据进行保护，以保证重要数据在清洗和转换前后的一致性，避免数据失真，并在出现异常时能有效还原和恢复被处理的数据；
e) 应跟踪和记录重要数据的采集、处理、分析和挖掘等过程，以通过溯源能重现相应过程；
f) 应采取物理破坏或使用无价值数据多次填充等手段，彻底销毁废弃存储介质上的数据。

6.7.1.2 数据生命周期管理

对数据生命周期的管理，宜采取“减少成本、减少风险”的策略，要求如下：

a) 应将数据的生命周期与存储级别相匹配，如活跃数据存放在在线存储中，非活跃数据存放在离线存储中；
b) 应积极管理数据的生命周期，主动管理数据的生命周期；
c) 应满足法律和审计要求；
d) 宜以减少信息管理风险为数据生命周期管理目标；
e) 宜以提高业务连续性为数据生命周期管理目标；
f) 宜以提高服务水平为数据生命周期管理目标。

6.7.2 元数据管理

元数据是描述数据的数据，与数据构造、数据流转、数据使用和数据维护密切相关，大数据平台应支持对元数据进行以下管理：

a) 应可以通过 SQL 脚本、API 等方式管理元数据；
b) 应可以通过手工编辑的方式管理元数据；
c) 宜使用 XML、EXCEL 存储表达元数据；
d) 应支持增加、删除和修改元数据；
e) 对于元数据的增量维护，应具备版本管理功能，如保留历史版本；
f) 应支持查询和统计元数据的使用情况。

6.7.3 系统资源管理

大数据平台应能集中管控大数据应用专属的计算和存储资源，要求如下：

a) 应支持按租户分配 CPU、内存、存储资源；
b) 应支持资源预留；
c) 应支持多级租户管理；
d) 应支持集群在线扩容或减容；
e) 应支持对辅助工具或服务组件进行管理；
f) 应支持屏蔽故障部分的计算、内存、存储资源。

7 基础资源子系统技术要求

应符合GB 50174—2017的要求。

8 安全运行子系统技术要求

大数据平台的网络安全等级保护管理应符合GY/T 337—2020的要求，安全运行应符合GB 50174—2017和GB/T 22239—2019的相关要求。

9 运维管理

9.1 概述

运维工作应包括运维能力体系建设、运维支撑保障基础建设、例行的各类运维操作和持续改进运维工作质量的过程管理等部分构成，通过对基础物理环境、数据资源、系统硬件、系统软件、应用软件、业务流程等对象的维护来对大数据平台的正常运转提供保障。

9.2 运维能力和支撑保障

要求如下。

a) 应组建运维团队，设置相应的部门，设置合理的岗位和人员管理机制；
b) 运维团队应具备能够及时发现系统故障或隐患的技术能力，装备能够及时发现系统故障或隐患、了解业务状态的监测、检测、监控工具，设置备品备件库。有条件的大数据平台运营单位，宜为运维队伍配置运维过程管理工具、资产管理工具、知识库等辅助工具。
c) 应编制运维服务对象和运维项目清单。
d) 应明确运维保障水平。
e) 应制定运维沟通协调机制。
f) 应规定运维考核方法。

9.3 运维操作

应对大数据平台各方面开展例行运维工作。

a) 对大数据平台的物理环境进行维护，要求如下：
   1) 应进行物理环境维护，指定专门的部门或人员负责机房管理，对机房出入进行管理；
   2) 应定期对机房供配电、空调、温湿度控制、消防等设施进行维护；
   3) 应按机房安全管理制度对物理访问、物品带进出和环境进行管理；
   4) 未经允许，不应在重要区域接待来访人员；
   5) 不应随意放置含有敏感信息的纸质文件和移动介质等。
b) 进行介质维护，要求如下：
   1) 应将介质存放在安全的环境中，对各类介质进行控制和保护，实行存储环境专人管理，并根据存档介质的目录清单定期盘点；
   2) 应对介质在物理传输过程中的人员选择、打包、交付等情况进行控制，并对介质的归档和查询等进行登记记录。
c) 进行设备维护，要求如下：
   1) 应对各种设备（包括备份和冗余设备）、线路等进行定期维护、维修；
   2) 重要数据处理设备应经过审批才能带离机房或办公地点，含有存储介质的设备带出工作环境时应对其中的重要数据进行加密；
   3) 存储介质或含存储介质的设备在报废或重用前，应进行数据完全清除。
d) 应对大数据平台进行日常监控，通过人工巡检或监控工具对大数据平台实施监控，获取系统的运行状态，及时响应大数据平台软件、硬件设备故障引发的业务中断或运行效率降低等引发的运维需求。
e) 应对大数据平台进行预防性检查，包括性能检查、脆弱性检查、漏洞扫描、恶意代码防范，如发现隐患及时进行评估处置。
f) 应对大数据平台开展常规运维作业，包括数据备份、配置备份、密码管理、系统升级、备件更换、日志分析、业务状态查询、业务流程人工干预等。
g) 应及时响应大数据平台运行需要、操作员或数据用户请求的运维需求。
h) 应有计划地对大数据平台的运行记录、趋势进行分析，并根据分析结果有针对性地改进、调整或升级大数据平台。
i) 应及时响应因各类原因引发的事件，如属应急事件，则应按照应急预案进行处置。

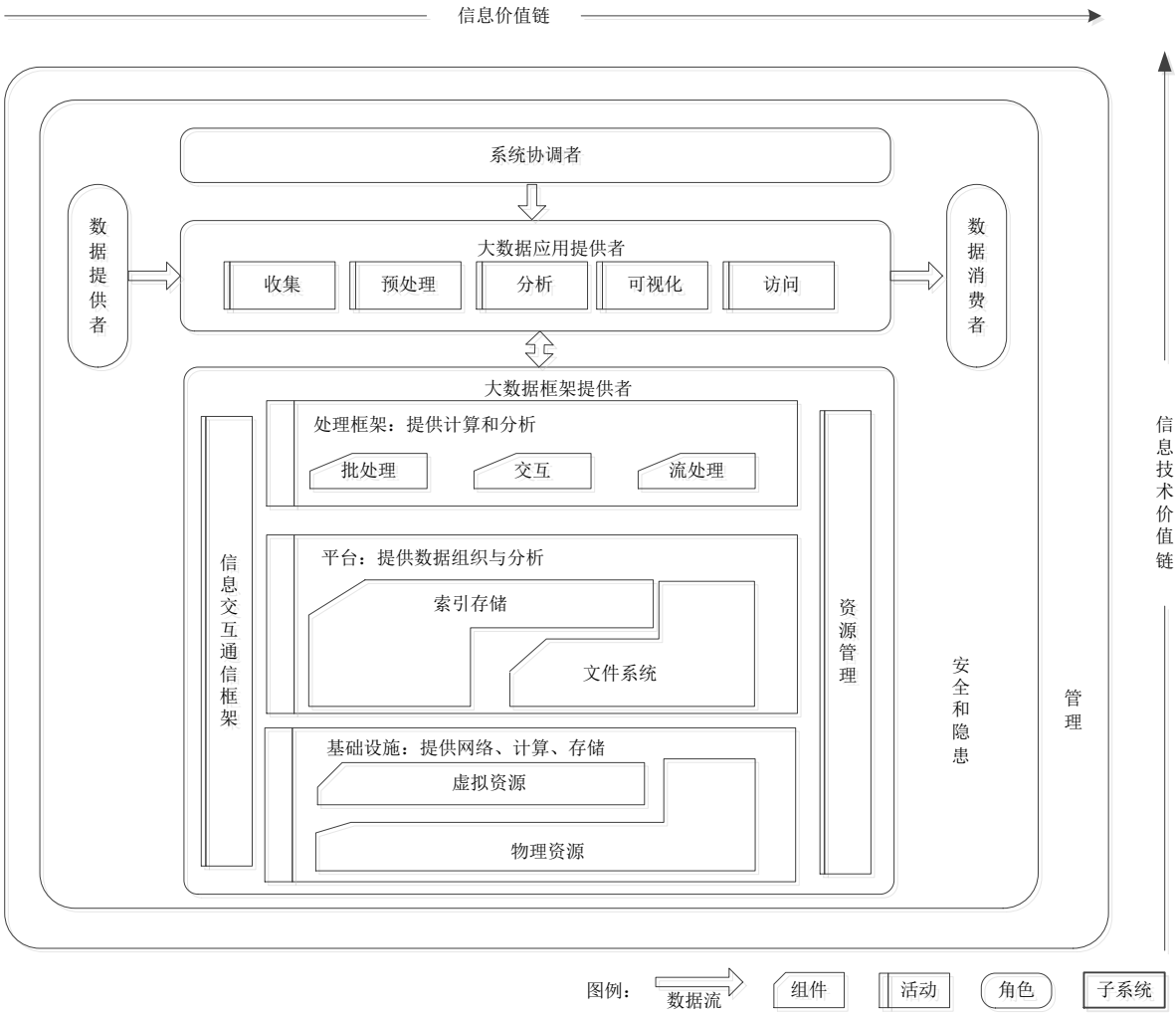
9.4 运维过程管理

要求如下：

a) 应开展运维项目清单管理，根据大数据平台改进、调整和升级情况，及时调整优化运维服务对象和运维项目；
b) 应开展数据资源管理，根据数字资源管理策略，对数据采集、存储、处理、应用、流动、销毁等过程进行管理；
c) 应开展运维保障水平管理，根据大数据平台改进、调整和升级情况，及时调整运维保障水平；
d) 应开展运维报告管理，对运维各类文档进行管理，例如做好巡检记录、总结报告、故障处理报告的撰写、审核、归档，对报告的准确性、完整性、及时性进行分析评估，不断提高运维报告质量；
e) 应开展事件管理，对运维过程中出现的事件受理、分类、分级、诊断、处置等过程进行分析，提高发现事件和解决事件的能力；
f) 应开展问题管理，识别事件发生的原因，预防或避免重新出现相同问题；
g) 应开展配置管理，对大数据平台的配置进行识别、记录，明确配置的审核、实施、回退、确认等操作过程，建立配置数据库，对配置变更进行管理；
h) 应开展变更管理，对大数据平台的变更请求、实施等过程进行评估、审核、确认和记录；
i) 应开展数据服务安全管理，对数据服务请求进行授权和审批，加强与内部人员和部门、各类职能机构、服务和设备供应商、业界专家及安全组织的合作与沟通，定期进行数据服务安全日常检查、汇总和通报；
j) 应开展应急管理，建立大数据平台应急管理机制，保证大数据服务在突发情况下通过采取应急手段得以继续运行；
k) 应定期进行信息安全等级测评，或在被保护系统发生重大变更或等级定级发生变化时进行等级测评。

附 录 A
（资料性）
大数据参考体系架构

GB/T 35589—2017对大数据标准体系进行了归纳总结，提出了图A.1所示的大数据参考体系架构。



图A.1 大数据参考体系架构

GB/T 35589—2017将大数据参考体系架构概括为“一个概念体系，二个价值链维度”。“一个概念体系”是指它为大数据参考体系架构中使用的概念提供了一个构件层级分类体系，即“角色—活动—功能组件”，用于描述参考架构中的逻辑构件及其关系；“二个价值链维度”分别为“IT价值链”和“信息价值链”，其中“IT价值链”反映的是大数据作为一种新兴的数据应用范式对IT技术产生的新需求所带来的价值，“信息价值链”反映的是大数据作为一种数据科学方法论对数据到知识的处理过程中所实现的信息流价值。

GB/T 35589—2017使用构件层级结构来表达大数据系统的高层概念和构件分类。从构成上看，大数据参考体系架构由一系列在不同概念层级上的逻辑构件组成，这些逻辑构件被划分为三个层级，从高到低依次为角色、活动和组件，其中：

——角色：处在构件的最顶层级，包括系统协调者、数据提供者、大数据应用提供者、大数据框架提供者、数据消费者、安全和隐私、管理；

——活动：处在构件的第二层级，是每个角色执行的活动；

——组件：处在构件的第三层级，是执行每个活动需要的功能组件。

本文件参照了GB/T 35589—2017大数据参考体系架构中使用的构件层级分类体系方法，以最大程度地表达大数据系统中不同的角色以不同的组件开展不同的活动这一主要思想。

参 考 文 献

[1] GB/T 22240—2008 信息安全技术 信息系统安全等级保护定级指南
[3] GB/T 25069—2010 信息安全技术 术语
[4] GB/T 35274—2017 信息安全技术 大数据服务安全能力要求
[5] GB/T 36073—2018 数据管理能力成熟度评估模型
[6] GA/T 1389—2017 信息安全技术 网络安全等级保护定级指南
[7] GY/T 317—2018 电视台信息系统运行维护服务通用要求
[8] T/31 SCTA001—2017 工业化大数据平台技术规范 数据采集接入
[9] T/31 SCTA002—2017 工业化大数据平台技术规范 数据存储
[10] T/31 SCTA003—2017 工业化大数据平台技术规范 数据处理
[11] T/31 SCTA004—2017 工业化大数据平台技术规范 数据展示
[12] GD/J 037—2011 广播电视播出相关信息系统安全等级保护定级指南
[13] GD/J 038—2011 广播电视播出相关信息系统安全等级保护基本要求
[14] 中国电子技术标准化研究院.大数据标准化白皮书（2020版）
[15] 全国信息安全标准化技术委员会.大数据安全标准化白皮书（2018版）
[16] 国家广播电视总局.广播电视和网络视听大数据标准化白皮书（2020版）