ICS 93.080.30
CCS P51

DB11

北　京　市　地　方　标　准

DB11/T XXXX—XXXX

人工智能算力中心建设规范

Standards for Construction of Artificial Intelligence Computing Center

××××－××－××发布　　××××－××－××实施

XXXX　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。

本文件由北京市经济和信息化局提出并归口。

本文件由北京市经济和信息化局组织实施。

本文件起草单位：北京电子控股有限责任公司、北京电子数智科技有限责任公司、北京电子商会、中国信息通信研究院、北京北方算力智联科技有限责任公司、北京信息基础设施建设股份有限公司、北京晶隆智算科技有限公司、中国移动通信集团北京有限公司、中国电信股份有限公司北京分公司、中国联合网络通信有限公司北京市分公司、中电云计算技术有限公司、京能数字产业有限公司、曙光数据基础设施创新技术 (北京)股份有限公司、阿里云计算有限公司、是石科技（平湖）有限公司、武汉元石智算科技有限公司、北京英赫世纪置业有限公司等。

本文件主要起草人：待定。

人工智能算力中心建设规范

1 范围

本文件规定了人工智能算力中心的技术要求，包括基础设施层、资源协调层、智算平台管理层、安全层等方面的要求。

本文件适用于新建人工智能算力中心的建设指引。

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 9813.3 计算机通用规范 第3部分：服务器

GB/T 42018 信息技术 人工智能 平台计算资源规范

GB 43630 塔式和机架式服务器能效限定值及能效等级

GB 50174 数据中心设计规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

人工智能算力中心 artificial intelligence computing center

面向人工智能应用，具备智能计算、存储、高性能网络、容器、安全等基础设施或服务，为各类智算场景和应用提供人工智能算力、大模型开发训练和统一监控运营等服务的数据中心。

3.2

数据中心 data center

为集中放置的电子信息设备提供运行环境的建筑场所，可以是一栋或几栋建筑物，也可以是一栋建筑物的一部分，包括主机房、辅助区、支持区和行政管理区等。

[来源：GB 50174，2.1.1]

3.3

算力资源 computility resources

具有计算能力、存储能力、网络能力等一种或多种能力的物理或虚拟资源。

[来源：GB/T 42018信息技术 人工智能 平台计算资源规范，3.4、3.5，有调整]

3.4

算力标识 computility identification

用于唯一标识算力资源的编码体系，包括城市、行业、企业、资源类型、数据中心、服务类型、计算、存储、功耗、网络等多个维度的信息。

3.5

算力调度 computility scheduling

是由调度系统依据特定策略和算法，感知计算任务需求和资源状态，动态决策、分配计算资源给任务，并协调其执行，以期优化资源利用率、提升任务性能、保障服务质量、降低运行成本的核心管理过程。

3.6

人工智能加速卡 artificial intelligence accelerating card

专为人工智能计算设计、符合人工智能服务器硬件接口的扩展加速设备，简称“加速卡”。

[来源：GB/T 42018，3.6]

注：文中的人工智能加速卡主要面向算力中心或服务器场景使用。

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）
CIFS：一种通用互联网文件系统协议（Common Internet File System）
CPU：中央处理器（Central Processing Unit）
CMIS：一种计算机管理接口规范（Common Management Interface Specification）
GPU：图形处理器（Graphics Processing Unit）
HDFS：一种分布式文件系统（Hadoop Distributed File System）
IPv4：互联网通信协议第四版（Internet Protocol version 4）
IPv6：互联网通信协议第六版（Internet Protocol version 6）
IT：互联网技术（Internet Technology）
NFS：一种网络文件系统协议（Network File System）
PUE：电能利用效率（Power Usage Effectiveness）
QPS：每秒查询率（Queries Per Second）
RDMA：远程直接内存访问（Remote Direct Memory Access）
RoCE：基于以太网的远程直接内存访问（RDMA over Converged Ethernet）
S3：简单存储服务（Simple Storage Service）

5 总体架构

人工智能算力中心可分为基础设施层、资源协同层、智算平台管理层和安全层，见图1。基础设施层为人工智能算力中心提供机房和IT基础设施等支撑；资源协同层对基础设施层的算力、资源进行统一调度和运营；智算平台管理层中，IT基础设施管理平台提供算力、存储、网络等服务，智算平台支持各类智算应用场景，数据服务平台支持平台全流程数据服务；安全层分为物理安全、网络安全、数据安全

和平台安全，保障人工智能算力中心基础设施、网络传输、数据资源及业务应用等多方面的机密性、完整性和可用性。

图1　人工智能算力中心总体架构图

6　基础设施层

6.1 机房基础设施

6.1.1 机房基础要求

建筑结构、供配电、暖通、布线系统、智能化系统、给水排水、机房环境、消防等方面应满足 GB 50174 有关要求。

6.1.2　节能要求

6.1.2.1　应采用高效能服务器等节能型设备。
6.1.2.2　应采用节能型制冷技术，当单机架功率大于20kW时，宜采用液冷技术方案。
6.1.2.3　新建项目PUE值应不超过1.25，且年能耗超过3万吨标煤的大规模项目PUE值应不超过1.15。
6.1.2.4　宜因地制宜利用氢能、太阳能等可再生能源资源。
6.1.2.5　宜采用余热回收技术。

6.1.3　其他要求

6.1.3.1　应具备便捷扩容维护能力。
6.1.3.2　供电、冷却、监控、物理安全和运维安全等方面基础设施或设备应具备高可靠性。

6.2 IT 基础设施

6.2.1　算力服务要求

6.2.1.1 塔式、机架式服务器应满足GB/T 9813.3和GB 43630相关要求。
6.2.1.2 服务器网卡插槽应有槽位物理标识，并与智能平台管理接口IPMI等系统读取内容一致。
6.2.1.3 服务器应支持多种供电节能模式，包含主备供电、关核、调频等功能，支持风冷或液冷散热。
6.2.1.4 服务器应支持供电的热插拔、液冷带液插拔。
6.2.1.5 风冷服务器应支持风扇模组热插拔，N+1冗余；液冷服务器应支持液冷盲插。
6.2.1.6 服务器应支持非系统硬盘热插拔。
6.2.1.7 人工智能训练服务器和人工智能推理服务器应满足GB/T 42018的要求。

6.2.2 存储服务要求

6.2.2.1 应满足AI训练全流程中原始数据存储、数据预处理、样本数据高速读取、断点续训、高速读写、模型文件可靠存储等需求。
6.2.2.2 应支持多种协议（S3、HDFS、NFS、CIFS）兼容互通。
6.2.2.3 应结合算力规模、大模型参数规模、输入输出模态等需求配置适当比例的全闪存储服务器和混闪存储服务器。
6.2.2.4 应支持接入InfiniBand或RoCE等无损网络。
6.2.2.5 应支持通过副本或纠删码（EC）方式，保证数据可靠性。
6.2.2.6 应支持全闪存、大容量盘、数据重删、数据压缩、增量快照等功能。

6.2.3 网络服务要求

6.2.3.1 智算中心网络结构

根据不同的业务功能，智算中心网络应分为核心网络、计算平面网络、存储平面网络、业务平面网络和管理平面网络，见图 2。

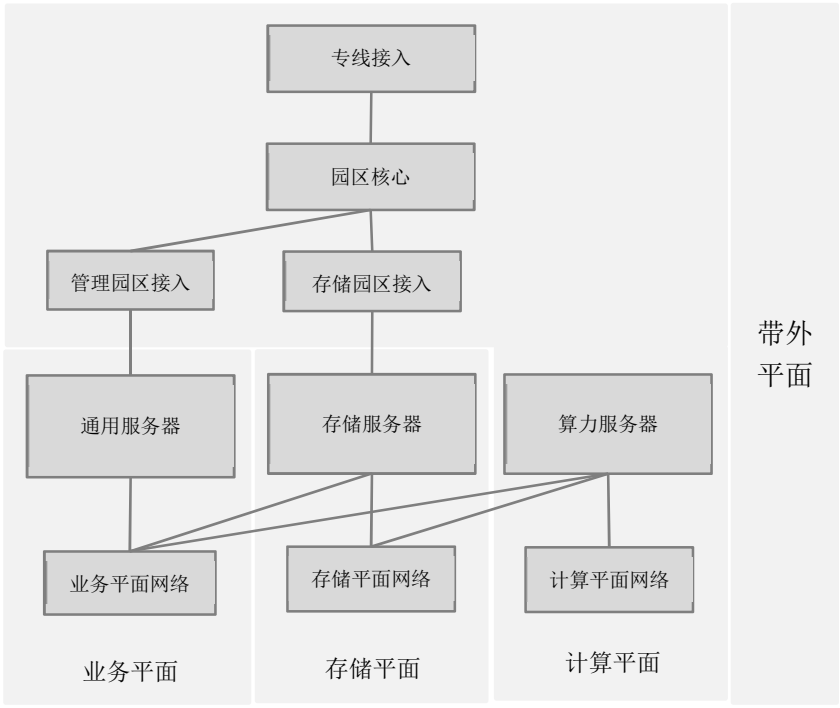

图2 智算中心网络结构图

6.2.3.2 各网络平面

6.2.3.2.1 网络应支持端到端 IPv4/IPv6双栈部署，并支持边界网关协议第四版BGPv4双栈路由协议；宜支持地址解析协议ARP或邻居发现协议ND表项与主机路由互通。
6.2.3.2.2 网络交换机应具备高可靠、高性能属性，满足大规模AI训练与推理的带宽和延迟要求；接入端口应满足对应平面对带宽速率的要求（≥10GE/25GE/200GE等）。
6.2.3.2.3 各网络平面应支持物理独立部署，确保不同业务流量间无冲突，保障训练性能。
6.2.3.2.4 各网络平面应具备租户隔离能力，支持逻辑隔离策略，确保不同业务或任务间的安全性和资源独立性。
6.2.3.2.5 网络设备宜采用Spine-Leaf或Core-Spine-Leaf架构组网，层级不超过三层，允许盒盒、框盒组网。应支持服务器、存储及管理设备的双归接入，以提升链路和设备的可靠性。各层级设备宜采用相同品牌、型号和批次产品。
6.2.3.2.6 光模块宜支持CMIS规范采集信息，包括产品标识、产品型号、生产商、序列号、生产日期、温度、收发光功率、光信噪比OSNR及纠前误码率BER等参数。

6.2.3.3 核心网络

6.2.3.3.1 应具备虚拟专用网络VPN划分和流量调度能力。
6.2.3.3.2 设备选型与容量设计应满足多人工智能算力中心园区对等互联组建广域网能力。
6.2.3.3.3 宜具备独立的互联网出口设备，宜具备独立的点到点专线接入设备。

6.2.3.4 计算平面网络

6.2.3.4.1 应支持作业间的逻辑隔离，以实现多任务并行训练场景下有效管理和资源分配。
6.2.3.4.2 应支持使用RoCE等高性能RDMA网络技术，支撑高性能训练与推理。
6.2.3.4.3 宜采用轨道优化方案，即各服务器同号网卡接入同号交换机或同号轨道组网，以兼顾并行计算拓扑亲和及更大规模，并由最上一级交换机进行合轨，支持服务器每张网卡双归接入。
6.2.3.4.4 宜支持智算服务器使用200GE以上速率端口接入支持RDMA网络。

6.2.3.5 存储平面网络

6.2.3.5.1 对有运营诉求的训练或推理集群宜支持保障数据安全，应支持安全隔离策略，包括计算节点间隔离、计算和存储节点间按需互访、未知节点缺省安全等隔离能力。
6.2.3.5.2 应支持使用RoCE等高性能RDMA网络技术，支撑高性能训练与推理。
6.2.3.5.3 宜支持存储服务器使用25GE以上速率端口接入网络，并支持智算服务器和存储服务器双归接入。

6.2.3.6 业务平面网络

6.2.3.6.1 宜与存储网络或带内管理网络融合组网。
6.2.3.6.2 宜提供10GE以上速率端口供业务服务器接入。

6.2.3.7 管理平面网络

6.2.3.7.1 管理平面网络应分为带内管理和带外管理两个互相隔离的网络，覆盖对算力设备、存储设备、网络设备的管理。

6.2.3.7.2 带内管理宜支持10GE速率以上端口，带外管理宜支持1GE速率以上端口。

7 资源协同层

7.1 算力资源

7.1.1 算力资源感知

7.1.1.1 应支持对计算、存储、网络多种不同资源类型的感知。
7.1.1.2 应支持对算力资源性能需求的感知，包括低时延、高移动性、大算力和潮汐性。
7.1.1.3 宜支持对算力资源应用场景需求的感知，包括生产需求、消费需求。
7.1.1.4 宜支持对不同维度算力资源的感知，如不同计算精度算力（包括但不限于FP64、FP32、FP16、INT8）与功耗水平（算效）。

7.1.2 算力数据统计

7.1.2.1 应支持对多种不同计算、存储、网络等算力数据的汇总统计。
7.1.2.2 应支持对不同维度算力数据的汇总统计。
7.1.2.3 应支持不同性能需求、应用场景需求下算力数据的汇总统计。
7.1.2.4 应支持按照场景、固定比例系数或者特定的计量单位进行具体的度量测算。

7.1.3 算力标识

7.1.3.1 应对异构物理算力资源建立统一的算力资源描述模型，根据不同需求与应用的算力指标，与该算力资源模型进行映射，以便对外部提供算力资源服务。
7.1.3.2 应适用通用算力、智能算力和超级算力等各类算力资源。
7.1.3.3 应支持对不同的算法如机器学习、神经网络等所需的算力。
7.1.3.4 应对数据中心算力资源提供统一的算力资源标识，便于上层应用的统一识别与调用。
7.1.3.5 应提供算力标识生成、标识解析、数据管理和服务接口等功能。

7.2 算力调度

7.2.1 算力调度能力

7.2.1.1 宜支持具备两种以上不同厂商、不同型号的人工智能加速卡算力的混合调度能力。
7.2.1.2 宜支持跨架构算力协同调度能力。

7.2.2 算力调度策略

7.2.2.1 应采用动态配置并按照配置顺序进行计算资源选择,组合使用多种调度策略,满足复杂的算力应用场景。
7.2.2.2 应支持负载感知调度，优先按照空闲计算资源匹配度高的集群进行作业调度。
7.2.2.3 宜支持任务调度和容器调度相结合的融合调度机制，要求如下：

a) 支持统一任务调度模型属性，例如队列、优先级、并行度等；
b) 支持通过 hook 等方式对外提供调度过程不同阶段执行的自定义流程；
c) 支持不同资源在不同任务执行之间的流动；
d) 支持多队列调度，并基于不同队列管理异构类型的计算资源，CPU、GPU、FPGA 等；

e) 支持多种任务执行类型，包括分布式任务、并行任务、大数据任务、AI 与深度学习任务；
f) 支持根据不同任务的资源需求确定资源分配策略，向资源管理模块申请相应资源；
g) 支持接收任务调度及容器调度对资源的申请，并在权限/配额允许的情况下，分配相应的资源；
h) 支持不同任务的运行状态的管理和查询。

7.2.3 算力集群调度适配规范

7.2.3.1 应开放区域协同管理接口，接口类型应包括验证鉴权、资源创建、资源配置、资源运行等生命周期管理，及监控采集。
7.2.3.2 应增加跨国家枢纽节点的调度接口，支持跨省域资源池化管理。
7.2.3.3 应通过合理分配和利用计算资源,提高计算机资源的利用率和任务执行效率同时屏蔽异构硬件和系统的差异，通过配置适配不同的目标集群。
7.2.3.4 应支持多种模型部署方式，包括裸机、容器等。

7.3 算力运营

7.3.1 应具备自服务能力，包括支持通过服务平台访问算力资源，支持平台咨询、试用、选购、售后等服务，支持进行算力、数据、软件等资源管理、监控与调整，支持进行组织、人员、对账等运营服务。
7.3.2 应具备按需服务能力，包括支持根据需求进行资源选择及使用，支持根据需求进行资源计量计费，支持服务平台可根据资源使用情况进行弹性伸缩。
7.3.3 应具备统一的用户及任务管理能力，包括具备统一的用户管理和用户环境配置、具备多种优先级的资源调度策略、具备多种策略管理能力，为作业请求提供最佳资源，支持数据分类分级授权。

8 智算平台管理层

8.1 IT 基础设施管理平台

8.1.1 算力纳管

8.1.1.1 应支持芯片、服务器、边缘节点、云平台等不同形式资源纳管。
8.1.1.2 应具备采集算力静态属性的能力，包括设备ID、型号、位置、硬件配置等信息。
8.1.1.3 应具备采集算力动态状态的能力，包括是否在线、开机/关机状态、健康状态等信息。
8.1.1.4 应支持算力资源的动态监控与弹性分配，支持容器自动识别和扩展。

8.1.2 网络服务

8.1.2.1 应提供高性能和低延迟的网络服务，支持超大规模网络，宜采用适合大模型训练网络流量特征的网络架构，如fat-tree等。
8.1.2.2 应支持RDMA高性能网络，如InfiniBand、RoCE，及相应的网卡、交换机。国产化场景应支持RoCE网络。

8.1.3 存储服务

8.1.3.1 应提供高性能的存储能力，如高性能存储、对象存储、块存储和文件存储等。
8.1.3.2 应提供并行文件存储。
8.1.3.3 应支持全闪高速存储和混闪容量存储方式。
8.1.3.4 应支持存算分离架构，存储系统支持对接多类计算引擎。

8.1.3.5 应支持断点续训场景。
8.1.3.6 应支持检索能力，提供较低的检索时延，针对百亿级数据宜提供ms级返回时延、单表PB级规模、支撑百列数据千万QPS低延迟写入能力。

8.2 智算平台

8.2.1 数据预处理

8.2.1.1 应支持特征工程，包括特征构建/提取、特征转换、特征选择、特征管理等。
8.2.1.2 应支持图像变换及数据增强、支持音频和视频的切分、转写。
8.2.1.3 应支持数据集协同、数据压缩和数据回流。
8.2.1.4 宜支持数据质量评估、数据版本控制、预训数据索引管理、数据质量问题反馈定位、支持数据随机抽样。

8.2.2 模型开发

8.2.2.1 应提供模型开发训练过程资源监控和日志管理。包括对任务的CPU、GPU、内存等资源的监控以及事件、训练日志的查看。
8.2.2.2 宜内置多种机器学习和深度学习算法样例模板和模型，用户可以开箱即用；用户也可以自定义算子或者模型发布到市场供其他用户使用。
8.2.2.3 宜内置多种主流编程语言和框架的镜像，支持用户自定义镜像的上传和使用，包括镜像环境的启动参数配置、镜像基本信息管理等。
8.2.2.4 宜支持多种训练方式，包括算子拖拉拽的可视化建模以及直接提供脚本镜像提交任务式建模等方式。
8.2.2.5 宜支持主流的开发工具链集成，用户可以在本地IDE进行开发。
8.2.2.6 宜支持自动建模。

8.2.3 模型训练

8.2.3.1 应支持多种分布式训练框架，支持数据并行、模型并行、张量并行等多种分布式并行。
8.2.3.2 宜支持CPU、GPU等异构计算和自动容错弹性训练，通过选择配置不同的异构资源池，对训练任务作业的弹性伸缩和容错处理，实现任务的高效调度。
8.2.3.3 应支持模型训练加速，包括数据并行训练、训练加速、自动显存优化等。
8.2.3.4 应提供可视化的分布式训练监控及日志管理。包括对任务的CPU、GPU、内存等资源的监控以及事件、训练日志的查看。
8.2.3.5 宜提供训练任务通过资源配额与命名空间方式隔离。

8.2.4 大模型微调

8.2.4.1 应支持大模型多种微调方式，应包含完整的微调工作流程，应支持指令调整、参数高效微调。
8.2.4.2 应支持开源模型与用户私有模型的微调。
8.2.4.3 应提供主流的微调框架，支持全量微调、高效微调、深度学习等。
8.2.4.4 应提供微调高质量数据集生成和数据抽取工具，应支持快速构建高质量微调数据集。
8.2.4.5 应提供提示词优化工具，为大语言模型提供自动化的提示工程。
8.2.4.6 应支持大模型知识增强，包括知识增广、支撑、约束和迁移。
8.2.4.7 应支持大模型与外界工具进行交互学习。

8.2.5 模型推理

8.2.5.1 应支持多种主流推理引擎，支持用户自定义推理服务框架等。
8.2.5.2 应支持灰度发布，支持对多个版本的模型文件设置不同的流量比例完成服务的稳定发布。
8.2.5.3 应支持查看推理任务的监控和日志，包括对任务的CPU、GPU、内存等资源的监控以及事件、训练日志的查看。
8.2.5.4 宜支持推理服务灰度发布的自动回滚机制，确保异常情况下的服务稳定性。
8.2.5.5 宜支持模型部署和自动扩展，用户可以通过设置QPS、请求并发数等指标阈值，实现服务的自动弹性伸缩。
8.2.5.6 宜支持服务编排。
8.2.5.7 宜支持在线和离线两种服务模式。
8.2.5.8 宜支持分组测试。
8.2.5.9 宜支持软硬件适配与协同优化，包括支持GPU、FPGA、ASIC等多种算力资源，解耦软硬件，实现大模型的高效推理。
8.2.5.10 宜支持大模型压缩，方法不限于模型稀疏化、权重矩阵分解、模型参数共享、蒸馏和量化。

8.2.6 模型优化

8.2.6.1 应支持自动化模型更新和升级。
8.2.6.2 应提供对算法训练调优过程的可视化工具，比如Tensorboard等用于查看、分析和监控训练过程。
8.2.6.3 应支持多种模型评估算法，比如混淆矩阵、二分类模型矩阵、回归、聚类等模型评估方法。
8.2.6.4 宜支持模型缓存和预热。
8.2.6.5 宜支持预测引擎扩展和升级。
8.2.6.6 宜提供多种算法、模型优化方法，通过特征工程对数据进行转换和加工，提取有效信息，提高算法的准确率；通过提供超参数优化算法，比如网络搜索、贝叶斯优化、遗产算法，提高算法、模型优化效率等。
8.2.6.7 宜支持多种模型互相转换，包括PyTorch、TensorFlow、Caffe等互转。

8.2.7 大模型多模纳管

8.2.7.1 应提供多规格、多版本、多类型的基础模型和任务模型的纳管，包括多种主流开源模型和闭源模型的APIs接入能力。
8.2.7.2 宜支持语言、知识、推理、多学科等模型评测方案，并支持大模型推理性能评测。
8.2.7.3 应支持边缘侧的统一纳管。
8.2.7.4 宜支持社区内分享模型、数据集和代码，促进领域内知识的传播与共享，共建大模型生态。

8.3 数据服务平台

8.3.1 数据汇聚与合成

8.3.1.1 应支持基于人工和大模型进行数据生成：支持基于主题、指令词与大语言模型APIs调用，快速生成高质量微调数据集。
8.3.1.2 应支持数据抽取，可以通过固定文档基于预设问题抽取数据（或答案），形成有效问答对形式，用于模型微调数据构建。
8.3.1.3 应支持对视频、图片、语音、文字等多模态的数据采集和处理。

8.3.1.4 宜支持通过人工智能技术生成高质量数据。
8.3.1.5 宜支持企业数据资产录入。
8.3.1.6 宜支持与数据登记平台对接。

8.3.2 数据标准化

8.3.2.1 应提供规范化的数据存储格式，支持对象存储和分布式存储。
8.3.2.2 应提供统一的数据接入方式，包括通信方式和数据报文格式。
8.3.2.3 应支持数据清洗方法以提升数据质量。
8.3.2.4 宜支持向量数据库。
8.3.2.5 宜提供数据质量评估工具。

8.3.3 数据标注

8.3.3.1 应提供面向人工智能应用的数据标注服务：包括对图像、文本、音频等多种数据类型的标注以及多模态的混合标注。
8.3.3.2 应提供丰富的标注内容组件以及预置的标注模板，支持多方协同进行数据的标注。

8.3.4 数据流通

8.3.4.1 应支持采用引入外部数据源进行模型训练、调优和推理。
8.3.4.2 宜支持多个外部数据源的接入。

8.3.5 数据产品开发

8.3.5.1 应对数据使用方和数据开发方对原始数据的使用过程采用技术方法进行安全控制，如基于身份的访问控制。
8.3.5.2 应严禁原始数据替代数据产品对外提供。
8.3.5.3 应支持通过API等至少一种接口方式为用户提供数据产品。
8.3.5.4 应支持对高价值高敏感数据产品生成数字水印。
8.3.5.5 宜支持大模型开发、数据统计分析等常见应用开发能力。

9 安全层

9.1 物理安全

9.1.1 安全防范系统

9.1.1.1 安全监控系统宜由视频安防监控系统、入侵报警系统和出入口控制系统组成，各系统之间应具备联动控制功能，视频监控系统应无盲区。
9.1.1.2 紧急情况时，出入口控制系统应能接受相关系统的联动控制信号，自动打开疏散通道上的门禁系统。
9.1.1.3 室外安装的安全防范系统设备应采取防雷电保护措施，电源线、信号线应采用屏蔽电缆，避雷装置和电缆屏蔽层应接地，且接地电阻不应大于10Ω。
9.1.1.4 安全防范系统宜采用数字式系统，支持远程监视功能。

9.1.2 环境及设备安全

9.1.2.1 应单独设置人员出入口和设备、材料出入口，通道宽度及门的尺寸不仅要满足设备和材料的运输要求，还应特别考虑高性能计算设备和液冷系统的特殊搬运和维护需求，确保设备安全运输和快速部署。
9.1.2.2 变形缝不宜穿过主机房，以确保复杂冷却系统和高性能计算设备的安全运行。
9.1.2.3 应具备远程视频监控系统和远程监控能力。
9.1.2.4 机房内应无爆炸性、导电性、导磁性及腐蚀性尘埃，确保设备安全。

9.2 网络安全

9.2.1 网络整体安全应至少满足等保三级要求。
9.2.2 应可分租户提供网络安全产品与服务。

9.3 数据安全

9.3.1 应提供基础防护、流量监控等安全服务。
9.3.2 应支持使用加密算法对敏感数据进行加密，包括数据传输过程中的加密（如超文本传输安全协议HTTPS）和数据存储时的加密（如数据库加密），以保证数据的机密性和完整性。
9.3.3 应支持防护和检测各种网络攻击。
9.3.4 应支持实时查看和处理安全事件，包括异常登录、攻击和病毒等。
9.3.5 应提供网络安全审计和日志分析。
9.3.6 应支持与数据源方之间建立安全通道，通过身份认证、数据加密等技术手段保证数据传输安全。
9.3.7 应支持对训练数据、输入提示词、检索增强生成RAG知识库以及关键模型参数的安全保护。
9.3.8 应支持算力中心间数据传输和消息通信的安全加密及认证。
9.3.9 应支持对外部敏感数据源的逻辑接入，保证敏感数据的使用安全。
9.3.10 宜采用数据沙盒、隐私计算、区块链、数字合约等可信管控技术保护数据隐私安全，防止数据滥用。
9.3.11 宜采用安全隔离措施针对不同的行业数据、不同用户数据进行分别保护。

9.4 平台安全

9.4.1 应支持严格的身份认证系统，确保只有经过授权的用户可以访问智算平台，同时根据不同用户角色和权限设置细粒度的访问控制策略。
9.4.2 宜支持基于角色的访问控制及多因素身份认证。

参 考 文 献

[1] GB/T 20272 信息安全技术 操作系统安全技术要求
[2] GB/T 22239 信息安全技术 网络安全等级保护基本要求
[3] GB/T 37939 信息安全技术 网络存储安全技术要求
[4] GB/T 39680 信息安全技术 服务器安全技术要求和测评准则
[5] GB 40879 数据中心能效限定值及能效等级
[6] GB/T 43331 互联网数据中心（IDC）技术和分级要求
[7] GB/T 45401.2 人工智能 计算设备调度与协同 第2部分：分布式计算框架