ICS 35.240.01
CCS L67

团 体 标 准

T/AI 126.3—2024

数据湖
第 3 部分：数据资源管理系统

Data lake applications
Part3：Data resource management system

2024－03－19 发布　　2024－03－19 实施

中 关 村 视 听 产 业 技 术 创 新 联 盟　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。

本文件是T/AI 126《数据湖》的第3部分。T/AI 126已经发布了以下部分：

——第 1 部分：磁光电混合媒体分布式存储系统；

——第 2 部分：蓝光存储资源管理系统接口；

——第 3 部分：数据资源管理系统；

——第 4 部分：人工智能技术应用要求；

——第 5 部分：城市治理水平评价模型；

——第 6 部分：交通应急指挥与协调决策系统接口。

本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由新一代人工智能产业技术创新战略联盟AI标准工作组提出。

本文件由中关村视听产业技术创新联盟归口。

本文件起草单位：北京易华录信息技术股份有限公司、文安智能科技有限公司、博雅鸿图视频技术有限公司、美的集团（上海）有限公司、北京安录国际技术有限公司、北京大学、中山大学。

本文件主要起草人：彭珂、王凌、李君、赵阳、倪志云、李冰青、谷桐宇、蔡亚森、汪志锋、黄铁军、赵海英、崔晓冉、李艳梅、梁凡。

引　言

数据湖是用于存储、处理和分析，大量结构化、半结构化和非结构化数据的系统或存储库。城市数据湖是融合数据感知、存储、分析为一体，以光磁融合存储为依托，以人工智能为引擎、以区块链、云计算、大数据平台等技术为支撑，提供IDC、云计算、湖存储、数据增值、数据安全等运营服务的新一代数字经济基础设施。

T/AI 126《数据湖》是指导城市数据湖建设和应用技术需求的基础性标准，数据湖应用技术涉及从数据分级存储、分析处理、AI需求定义、应用技术框架、行业应用业务落地等全方位覆盖产业链各环节。基于此，本标准从多维多源数据的入湖、在湖、出湖各场景应用技术需求给出标准系列，拟由六个部分构成。

——第 1 部分：磁光电混合媒体分布式存储系统。数据湖中汇集了海量数据，磁光电混合媒体分布式存储系统作用于数据入湖阶段，目的在于指导数据湖中磁光电混合存储的应用系统设计、建设和应用，提供技术参考。

——第 2 部分：蓝光存储资源管理系统接口。目的在于指导蓝光光盘存储的应用系统设计与开发，提供技术参考。

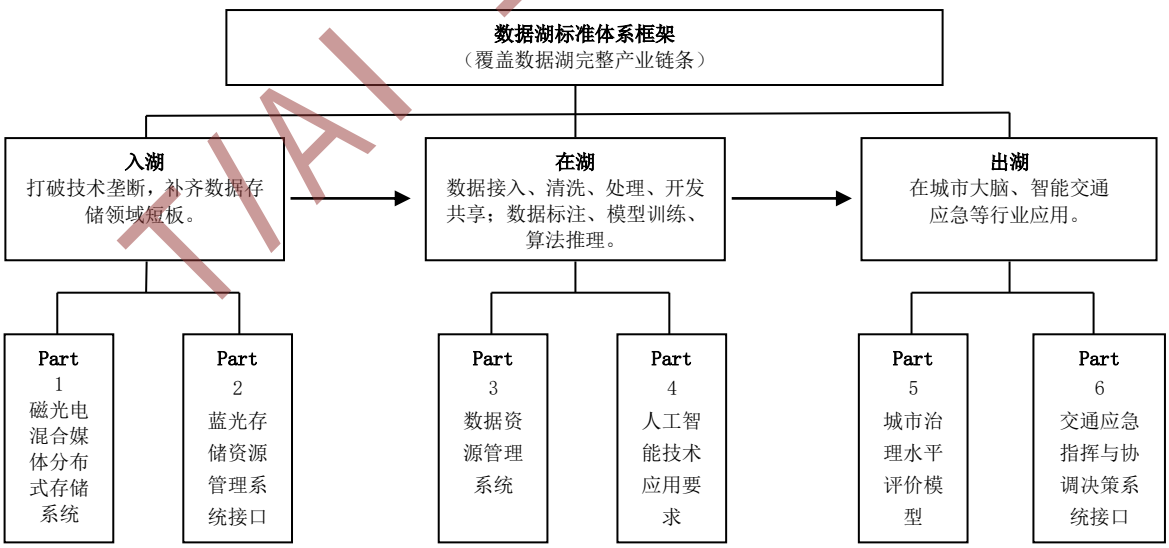
——第 3 部分：数据资源管理系统。目的在于针对数据湖内数据的接入、存储、管理、共享交换等，提供技术参考。

——第 4 部分：人工智能技术应用要求。目的在于指导数据湖中人工智能技术应用框架及人工智能技术应用 API 的设计与开发，提供技术参考。

——第 5 部分：城市治理水平评价模型。目的在于针对数据湖城市大脑建设过程中对指标的统一规划、统一开发、统一运维等做一定程度的指导，提出技术参考建议。

——第 6 部分：交通应急指挥与协调决策系统接口。目的在于明确定义了基于数据湖支撑的交通应急决策系统相关接口技术，为面向城市区域交通安全事件应急指挥与协调决策提供相关系统建设提供技术参考。

数据湖标准体系框架如下所示：



数据湖
第3部分：数据资源管理系统

1 范围

本文件给出了数据资源管理系统的功能架构，规定了系统的基础支撑、数据集成、数据管理、数据交换、数据分析挖掘、数据备份、数据安全和SaaS软件运营服务等要求。

本文件适用于部署在数据湖上的数据接入、标准化清洗、管理、处理、开发共享等应用，可用于指导数据治理类工具产品的设计与开发。

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。

GA/T 1718-2020 信息安全技术 大数据平台安全管理产品安全技术要求

GB/T 37973-2019 信息安全技术 大数据安全管理指南

GB/T 35273-2020 信息安全技术—个人信息安全规范

YD/T 3595.1-2019 大数据管理技术要求 第1部分：管理框架

3 术语和定义

下列术语和定义适用于本文件。

3.1

数据资产 data asset

合法拥有或者控制的，能进行计量的，为组织带来经济和社会价值的数据资源。

3.2

数据集成 data aggregation

指从多个来源收集数据，以便将所有数据放到一个共同的资源池中，以进行报告和/或分析。

3.3

数据仓库 data warehouse

在数据准备之后用于永久性存储数据的数据库。

[来源：GB/T 35295-2017，定义2.1.35]

3.4

多租户 multi-tenancy

实现在多用户（一般指面向企业用户）环境下共用相同的系统或程序组件，并且可确保各用户间数据的隔离性。

3.5

分布式系统基础架构 the architecture of distributed systems

是一个能够对大量数据进行分布式处理的软件框架。

3.6

元数据 metadata

关于数据或数据元素的数据（可能包括其数据描述），以及关于数据拥有权、存取路径、访问权和数据易变性的数据。

[来源：GB/T 36073-2018，定义3.8]

4 缩略语

下列缩略语适用于本文件。

API：应用编程接口（Application Programming Interface）
ES：弹性搜索（Elastic Search）
ETL：提取、转换和加载（Extract-Transform-Load）
IDE：集成开发环境（Integrated Development Environment）
KaaS：托管基础设施即服务（Kubernetes as a Service）
K8s：用于管理云平台中多个主机上的容器化的应用（kubernetes）
MPP：大规模并行分析数据库（Massively Parallel Processing）
MR：用于大规模数据集并行运算的编程模式（MapReduce）
OCR：光学字符识别（Optical Character Recognition）
RBAC：基于角色的访问控制（Role-Based Access Control）
SaaS：软件运营服务（Software as a Service）
SQL：结构化查询语言（Structured Query Language）

5 系统功能架构

数据湖数据资源管理系统应由8个部分组成，具体包括：基础支撑、数据集成、数据管理、数据交换、数据分析挖掘、数据备份、数据安全、SaaS。数据资源管理系统功能架构见图1。

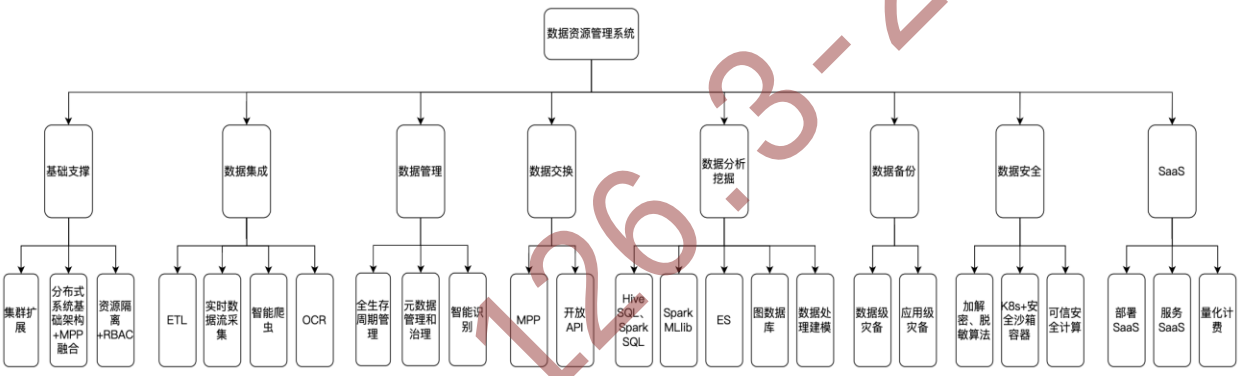


图1 功能架构

6 基础支撑

6.1 概述

数据资源管理的底层基础架构，为业务系统开发提供一个涉及项目整个生存周期的快速集成环境。应包括：集群扩展、分布式系统基础架构+MPP、资源隔离+RBAC等技术。

6.2 集群扩展

6.2.1 一键部署接口调用

为便于安装大数据平台，可采用通过页面操作自定义安装组件，实现在线扩容，使其具备横向扩展能力。

6.2.2 集群联邦技术应用

为避免区域连接中断或各类故障导致任务分配失败，集群扩展应采用联邦技术，在服务器后台中，按照地理位置创建一个复制机制，将多个集群进行复制，或者跨集群冗余部署。故障时，任务重新分配给集群联邦中其他可用状态的集群上，保障最关键的服务运行。

6.3 分布式系统基础架构+MPP 融合

数据湖为满足海量多源异构数据的多场景应用需求，基础支撑架构应采用分布式系统基础架构+MPP混搭架构，提供统一SQL编程，屏蔽异构数据源访问差异，实现多源异构数据融合分析。

6.4 资源隔离+RBAC

为提高数据资源管理效率和数据使用安全性，应用层数据管理工具宜采用资源隔离技术，在多租户集群中，对不同的租户应提供尽可能的安全隔离，防止恶意租户对其他租户的攻击，保证租户之间公平地分配共享集群资源。资源隔离应采用以RBAC为核心技术的隔离方式，引入角色和权限进行解耦，简化授权和安全管理。

7 数据集成

7.1 ETL

数据湖数据集成宜采用ETL工具，通过工具提供的数据管道，可以将数据从内部存储库集中转储到基于云的存储中，实现数据库的组织和结构（包括表、列、数据类型、视图、存储过程、关系、主键和外键等）快速感知变化。

7.2 实时数据流采集

针对流数据，宜采用Flume实时数据流采集框架，通过编写一个配置文件，定义读数据的源端，数据写入的目标端，运行配置文件，数据一旦产生，可以被立即采集，保证准实时性。

7.3 智能爬虫

针对网页数据，宜采用智能语义爬虫，根据语义识别，自动进行信息格式化分析，非定向抓取系统靶向信息，有效地保证信息数量，同时不应使用人工参与定制和维护模板，有效地保证了自身的人力和维护成本。

7.4 OCR

针对文本数据，宜采用文档识别技术功能特点图像输入、图像预处理、版面分析、字符切割、字符特征提取、字符识别、版面恢复、后处理校正等算法。

8 数据管理

8.1 全生存周期管理

针对结构化、视频、图片及文本各类格式数据，应采用涉及数据接入、清洗转化、归档、销毁数据、分级分类管理的全生存周期管理，实现存储策略的灵活设置，数据的稳定高速读取。

8.2 元数据管理和治理

数据湖数据管理可采用元数据管理和治理，以及符合YD/T 3595.1-2019的元数据管理相关要求，实现管理共享元数据、数据分级、审计、安全性以及数据保护等方面功能，提升数据分析师和数据治理团队对企业数据资产进行分类和管理的能力。

8.3 智能识别

智能识别可实现便捷的清洗和关联分析，元数据和敏感数据智能识别，数据资产实现自动分级分类，把关数据质量，生成标签，形成全局统一的数据视图。

9 数据交换

9.1 MPP

数据湖数据资源管理系统可采用MPP等兼具可伸缩性、高可用、高性能、优性价比等优势的工具，替代现有关系型数据库，以提升大数据处理，及数据资源共享效率。

9.2 开放API

共享交换宜采用开放API与RBAC技术融合的方式，实现系统之间数据交流的有效隔离，即开发者用户无需访问源代码，就可依据权限进行文档或网络流量检查，实现逻辑最小化来理解远程服务并与之交换，有效控制数据访问权限的，达到保障安全的目的。

10 数据分析挖掘

10.1 Hive SQL，Spark SQL

数据分析挖掘，可采用Hive SQL，Spark SQL，对结构化、半结构化等各类数据源实现数据的脚本开发利用、数仓构建、指标数据的输出、API的生成。

10.2 Spark MLlib

数据分析挖掘，可通过Spark MLlib中封装好的算法，如：分类、回归、聚类、协同过滤、降维以及底层的优化原语等算法和工具，快速构建业务分析模型，将多种算法组合到单个工作流或pipeline中，可支撑数据资源管理系统中逻辑性强的分析挖掘场景。

10.3 ES

数据分析挖掘，可采用ES等工具在全文字段中搜索，对主表数据进行快速的分词和查询定位。

10.4 图数据库

数据分析挖掘，可采用图数据库，基于“节点（实体）+边（关系）”的架构，将结构化表格数据转化为“节点”与“边”的图谱关系图，实现可视化配置。构建业务图谱，应支撑千亿节点（千亿条关系图数据计算场景），快速进行数据的查询分析和关系展现。

10.5 数据处理建模

数据分析挖掘，可采用数据处理建模工具，实现数据上传、数据集成、跨项目发布、快速部署和协同开发等，以提升开发工作效率。

a) 提供本地数据上传功能，实现本地文本数据实现云存储；
b) 提供可视化工作流程设计器功能，对流程进行设计并编辑，对流程中的每一个任务节点进行相应的开发工作；
c) 提供海量异构数据源的数据快速集成能力；
d) 提供 Web IDE 编程和调试环境，可使用 SQL、MR、数据同步等多种程序类型；
e) 提供跨项目发布能力，快速将任务及代码部署到其他工作空间的调度系统；
f) 提供协同开发，实现代码版本管理，多人协同模式下的代码锁管理和冲突检测机制。

11 数据备份

11.1 概述

数据湖数据资源管理系统，为PB级的数据集提供分布式数据库备份服务，实现单网络、跨网部署模式，同时实现数据湖间数据同步，平台级联。宜采用数据级灾备和应用级灾备，以保证在灾难发生后能够快速、准确的恢复客户的业务数据和关键应用系统，保障客户业务的持续运行。

11.2 数据级灾备

数据级灾备依靠基于网络的数据复制工具，实现生产中心和灾备中心之间的异步/同步数据传输，确保客户的原有数据不被破坏。

11.3 应用级灾备

应用级灾备应具备应用系统接管能力，即在异地灾备中心再构建一套支撑系统、备用网络系统等部分。生产环境发生故障后，灾备中心可以接管应用继续运行，减少系统宕机时间，保证业务连续性。

12 数据安全

12.1 概述

数据安全模块应当具备加解密、脱敏算法，K8s+安全沙箱容器，可信安全计算等多个技术手段，从传输、存储、计算等多方面全方位保证数据的安全性。

12.2 加解密、脱敏算法

针对密级数据，如私密或者绝密数据等级别的数据，应遵从法律法规和政府、企业机构的要求，以及符合GB/T 37973-2019、GB/T 35273-2020、GA/T 1718-2020数据安全的相关要求，应采用数据加解密、数据脱敏的措施在用户进行传输的过程中保障数据安全，维护个人、企业和政府的隐私。

12.3 K8s+安全沙箱容器

12.3.1 企业内部共享集群的多租户

根据企业内部人员结构的复杂程度，应采用K8s集群对不同部门或团队进行资源的逻辑隔离，在不同的命名空间之间只能够允许白名单范围内的跨租户应用请求。

12.3.2 多企业 SaaS & KaaS 服务模型下的多租户

在多个企业SaaS & KaaS场景下，为保证多租户间网络和资源配额上的隔离，宜采用安全沙箱容器来实现容器内核级别的隔离，才会最大限度避免恶意租户通过逃逸手段的跨租户攻击。

12.4 可信安全计算

12.4.1 区块链

数据资源管理系统，可采用区块链技术，对不同部分的数据流通记录进行存证，确保各个环节数据安全可溯。

12.4.2 联邦学习

数据资源管理系统，可采用联邦学习（在保证数据隐私安全及合法合规的基础上，实现共同建模，提升AI模型的效果）等基于多方安全计算的技术，通过私有化+SaaS部署方式，实现联邦分析、数据脱敏等能力保障数据安全计算。

12.4.3 xID

数据资源管理系统，可采用xID等技术，在不复原原始标识的条件下，将归属于同一数据主体的不同xID标记进行重标识，实现可控的数据流通。

13 软件运营服务 SaaS

13.1 概述

数据资源管理系统，在部署、服务两方面均应采用SaaS技术，实现资源优化配置，另外，还应采用控制SaaS平台的计费方式。

13.2 SaaS 部署

部署SaaS，应采用K8s+docker+云存储+云计算技术，对安全、网络、存储等做基础配置，对接数据湖云存储，构建统一的数据资源池，按需使用、弹性伸缩，计算存储分离。实现全平台快速部署，降低运维难度，提升资源使用效率，

13.3 SaaS 服务

服务SaaS可采用数据仓库、数据湖两类技术，内部打通引擎间差异，通过极简配置和缓存技术，简化上云工作量，降低用户业务升级难度。

13.4 量化计费

13.4.1 概述

在SaaS部署和服务后，可采用量化服务实现运营收费，推荐资源计费、API调用次数计费和用户数量计费三种计费方式。

13.4.2 资源计费

采用资源计费，包括计算资源量和存储资源量二种计算方法，所有被提交至调度系统的节点所产生的各类型实例（虚拟节点产生的实例除外），均可以做计算资源量来计费；所有存储至大数据平台的数据容量可以做存储资源量来计费。

13.4.3 API 调用次数计费

采用API调用次数计费。

13.4.4 用户数量计费

采用用户数量计费，通过计算用户的注册数量来计费。

参 考 文 献

[1]GB/T 35295-2017　信息技术 大数据 术语
[2]GB/T 36073-2018　数据管理能力成熟度评估模型