



华 西 证 券 股 份 有 限 公 司 企 业 标 准

Q/HXZQ 002—2025

华西证券数据平台数据入湖技术规范

Technical Specification For Data Lake Ingestion of HUAXI Data Platform

2025-12-01 发布

2025-12-01 实施

华 西 证 券 股 份 有 限 公 司 发 布

目 次

前言 II

引言 III

1 范围 1

2 规范性引用文件 1

3 术语和定义 1

4 数据分类 1

5 入湖原则 4

6 入湖方式 4

7 入湖流程 6

附录 A（资料性） 数据入湖要素及验收要素 8

 A.1 数据入湖要素 8

 A.2 数据验收要素 8

参 考 文 献 10

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范》的规定起草。

本文件由华西证券股份有限公司提出。

本文件由华西证券股份有限公司归口。

本文件起草单位：华西证券股份有限公司信息技术部。

本文件主要起草人：巩喜云、陈薇、刘光强、戢涛、常浩漾、余小平。

本文件及其所代替文件的历次版本发布情况为：

——2025 年首次发布为，Q/HXZQ 002—2025；

——本次为首次发布。

引 言

湖仓一体化大数据中心，作为数据基础设施，使用分布式与大数据技术承载全公司数据。数据湖的目标是存储所有公司数据，用于建仓的ODS层数据是其子集。为规范数据入湖相关的技术标准，结合公司信息系统现状，制定本技术规范，用于指导数据入湖工作。

本文件的制定依据国家相关法律法规及行业标准，结合实际业务与技术需求，提供科学、可操作的技术指导，适用于华西证券数据平台建设工作。

华西证券数据平台数据入湖技术规范

1 范围

本文件将公司数据资产进行分类，制定入湖原则、入湖方式、入湖流程等技术规范，为数据入湖整体工作提供指导。

本文件适用于指导华西证券内部开展数据入湖工作过程中的管理规范。

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。

JR/T 0304.1—2024 证券期货业基础数据元规范第1部分：基础数据元

Q/HXZQ 001—2024 华西证券湖仓一体化大数据中心数据存储技术规范

3 术语和定义

下列术语和定义适用于本文件。

3.1

湖仓一体化大数据中心 HUAXI Data Lakehouse

基于大数据技术架构，实现湖仓一体化大数据中心，提供数据的统一存储、计算和管理功能，满足全公司的数据需求。

3.2

数据入湖 Data Lake Ingestion

从业务系统或其他文件介质，采集数据，集成到湖仓一体化大数据中心的过程。

4 数据分类

截止2025年底，公司业务数据全景如下：

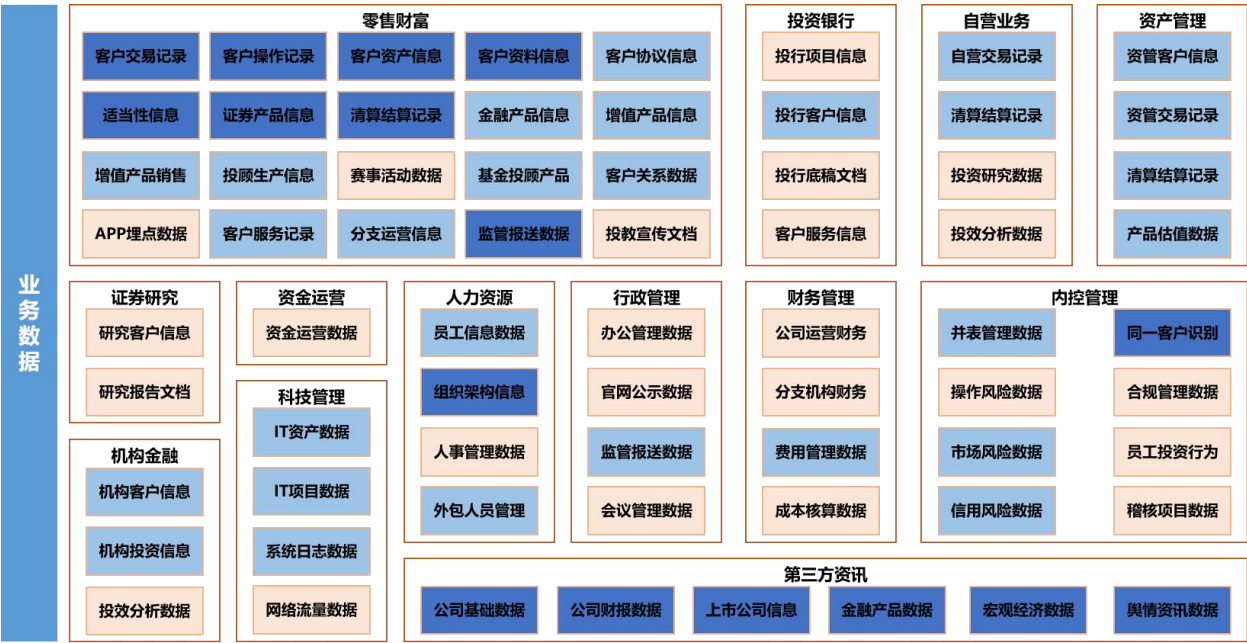


图 1 公司业务数据全景

数据分类如下：

表 1 数据分类表

存储方式	数据来源	分类名称	说明	入湖方式
数据库	在线系统	在线关系型数据	使用关系型数据库存储，由在线业务系统产生，数据持续变更	常规 ETL 核心数据切片 实时数据采集
		在线非关系型数据	使用非关系型数据库（如图数据库）存储，由在线业务系统产生，数据持续变更	专项数据采集
	已下线系统	离线关系型数据	已下线系统的备份数据库，数据稳定无变更	常规 ETL
流式	系统接口	流式接口数据	业务系统 API 接口数据、Kafka 订阅埋点数据、安全设备 syslog 等方式提供的流式数据	接口采集
文件	日志文件	日志文件	在线系统（业务系统、操作系统）产生的日志文件、系统性能文件	日志采集
	数据库备份	数据库备份文件	静态的数据库备份文件	数据库恢复采集

	人工台账	人工填报数据	无系统承载的台账数据，由人工维护	数据填报
	其他文件	其他文件	其他文档、音视频文件、清算文件等文件资料	文件对象存储

根据数据的存储方式，分为数据库、流式、文件三类，结合数据来源，可细分以下为8类：在线关系型数据、在线非关系型数据、离线关系型数据、流式接口数据、日志文件、数据库备份文件、人工填报数据、其他文件。

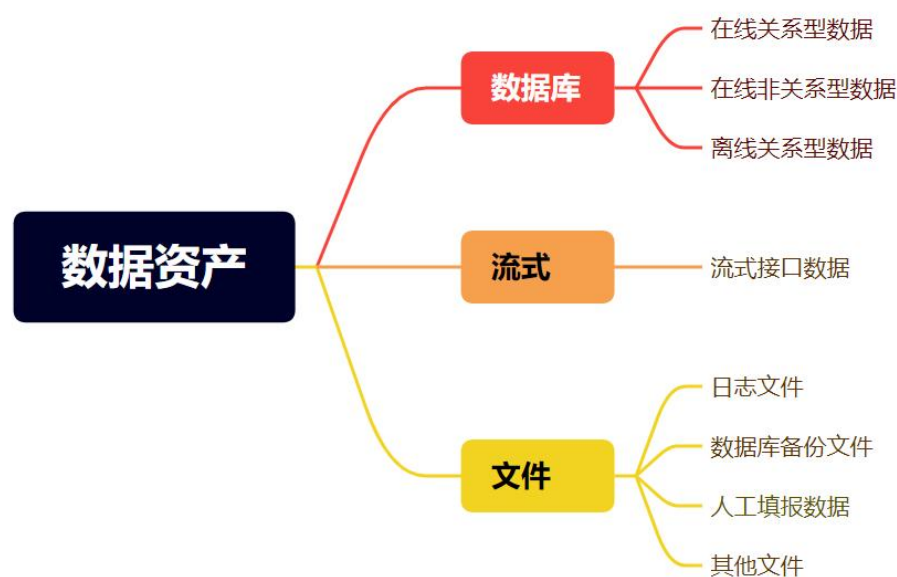


图2 数据资产分类

4.1 数据库

- 在线关系型数据：应用于在线业务系统，存储于关系型数据库，如交易系统客户交易数据、账户系统客户资料信息。根据下游数据应用的需求，分别使用常规 ETL、核心数据切片、实时数据采集等技术入湖。
- 在线非关系型数据：应用于在线业务系统，存储于非关系型数据库，如 CMDB 系统保存于图数据库的实体关系信息。对于非关系型数据库，需采用专项技术进行采集对接。
- 离线关系型数据：未应用于在线业务系统的数据库，通常为已下线系统的备份数据库（目前均为关系型数据库），数据稳定无变更，如已下线交易系统 kd32 的远期历史数据。此类数据采用传统 ETL 进行一次性采集入湖。

4.2 流式

- 流式接口数据：未落地，以接口方式提供的流式数据，如业务系统 API 接口数据、Kafka 订阅埋点数据、安全设备 syslog 等方式提供的流式数据。此类数据采用接口采集方式进行对接。

4.3 文件

- 日志文件：由业务系统、操作系统等产生运行日志文件，如手机 APP 运行日志、交易系统中间件功能号统计日志等。此类数据使用日志采集工具（统一日志管理分析系统）进行采集解析，将分析结果入库。
- 数据库备份文件：由在线系统备份产生的静态数据库文件，如集中交易系统每日清算备份文件。此类数据需将备份文件恢复至数据库中再进行 ETL 采集。
- 人工填报数据：非业务系统产生的数据，由人工维护，如监管报表涉及非系统生成的数据、数据报表相关输入数据。此类数据通过统一的数据填报入口，需经人工填报、审核后入湖。
- 其他文件类：离线的非结构化数据，如客户影像资料、回访短信语音记录、中登交易所及银行清算文件等。此类数据仅使用对象存储技术保存，暂不做解析，后续结合大模型等进行解析。

5 入湖原则

5.1 原始记录

在数据湖层面，不对数据做任何转换、清洗、加工，所有表均全字段采集；对于不同平台技术存储数据结构的差异，应尽可能进行准确的结构映射，如确实存在数字精度不明确问题，可用字符串进行过渡。

5.2 来源清晰

数据湖的存储命名规范，参考《湖仓一体化大数据中心数据存储技术规范》对ODS层数据的相关规范；根据入湖数据库表命名规则等规范，可明确数据来源的业务系统、库、表。

5.3 规则明确

入湖的依赖条件、采集时点频次、落地形态、保存期限等技术标准需明确，执行规则明确并保持一致性。

5.4 权限可控

作为公司级数据平台，数据湖应包含尽可能多的有价值数据，因此在采集业务系统数据时相对宽松（除公募基金客户等特定客群数据、员工职级薪资等数据外，按业务系统数据库粒度授权），但对外的数据输出或接口访问需严格按照最小化原则提供数据。

5.5 类型多样

数据湖使用多种类型的技术方案，支持不同的入湖数据类型，满足不同数据的存储和使用需求。

6 入湖方式

对于不同的数据分类，使用相应的入湖方式。

6.1 常规 ETL

- a) 采集技术：传统ETL，通过数据库连接在源业务系统数据库中查询获取需采集数据，并批量写入数据湖。
- b) 依赖条件：定时+数据就绪标识（可选）。

- c) 采集策略：数据采集频次（年/月/自然日/交易日/若干小时/若干分钟）；采集类型（全量/增量）；历史数据处理方式（是否需一次性补采）。
- d) 落地形态：参考源系统表数据结构，增加必要的系统标识字段（如入库时间），在数据湖中建立对应表进行存储。

6.2 核心数据切片

为了解决传统ETL采集数据有时间差，导致在实时变化的数据可能产生表间数据不一致问题，在交易系统清算后数据固定无变化的时间点，通过BCP方式获取指定表的切片数据，用于客户资产计算等核心应用场景。

- a) 采集技术：在指定运维环节中，通过BCP/DUMP方式将指定表（新增采集时间字段）从核心交易系统导出为CSV格式，数据平台对文本文件进行解析入湖。
- b) 依赖条件：清算完成，外部数据报送无异常反馈；数据BCP导出就绪后生成指定文件进行标识；需考虑不同核心清算进度差异的问题。
- c) 采集策略：数据采集频次固定为交易日；采集类型固定为全量。
- d) 落地形态：参考源系统表数据结构，增加必要的系统标识字段（入库时间），在数据湖中建立对应表进行存储。

6.3 实时数据采集

- a) 采集技术：通过解析数据库日志（需源端数据库开放，如Oracle需借助OGG等工具），建立目标数据的更新机制，实现实时数据同步。
- b) 依赖条件：无。
- c) 采集策略：实时采集。
- d) 落地形态：与源系统表结构一致，并可进行匹配的数据操作。

6.4 专项数据采集

- a) 采集技术：需使用非关系型数据库特定技术进行采集，针对特定类型确定技术方案。
- b) 依赖条件：定时+数据就绪标识（可选）。
- c) 采集策略：数据采集频次（年/月/自然日/交易日/若干小时/若干分钟）；采集类型（全量/增量）。
- d) 落地形态：根据特点类型，确定以匹配的相似组件存储，或转换为通用格式存储。

6.5 接口采集

- a) 采集技术：调用业务系统API接口，或订阅Kafka、syslog消息流数据，数据平台解析后入湖。
- b) 依赖条件：接口调用。
- c) 采集策略：实时采集。
- d) 落地形态：已转换为结构化数据的，在数据湖中建立对应表进行存储；未转换的以字符串格式保存；或以数据流方式对外提供服务。

6.6 日志采集

- a) 采集技术：通过日志系统采集原始文件进行解析；结构化后可保存到数据库，通过ETL方式进行采集入湖，如华彩客户登录记录。
- b) 依赖条件：通过日志系统完成采集并结构化输出。
- c) 采集策略：增量。

d) 落地形态：在数据湖中建立对应表进行存储。

6.7 数据库恢复采集

- a) 采集技术：通过脚本辅助，实现数据库文件恢复、ETL采集指定范围表并批量写入数据湖。
- b) 依赖条件：明确ETL采集范围（交易系统参考核心数据切片范围），验证数据库恢复效果。
- c) 采集策略：指定采集范围全量。
- d) 落地形态：参考源系统表数据结构，增加必要的系统标识字段（如入库时间），在数据湖中建立对应表进行存储。

6.8 数据填报

- a) 采集技术：通过数据平台统一的数据填报功能，填报人员录入或导入Excel文件，审批人员审批通过后，相关数据保存到数据湖。
- b) 依赖条件：数据通过审批。
- c) 采集策略：审批通过后触发。
- d) 落地形态：根据录入或导入文件格式建表，保存相关数据（附带日期或版本信息）。

6.9 文件对象存储

其他文件类，前期仅以数据保存为目标，后续考虑结合大模型等技术提取图像、音频、视频中有效的信息。

- a) 采集技术：文件读取，以对象存储方式进行文件存储。
 - b) 依赖条件：建立对象存储集群技术体系。
 - c) 采集策略：一次性全量。
- 落地形态：对象存储集群保存。

基于分布式的数据分片与副本复制技术，避免单点故障导致的数据丢失，保障数据存储安全。

7 入湖流程

7.1 需求管理

a) 需求类型

主动需求：基于规划的数据范围，以建立公司级数据平台为目标，梳理数据来源的业务系统，主动接入数据。

被动需求：基于平台上承载项目的需求，或数据同步需求，扩展接入业务系统的范围。

b) 权限管理

常规数据：以库为单位申请权限，按需分批次接入数据表。

特殊数据：对于公募基金客户等特定客群数据、员工职级薪资等特殊数据，需部门及关联业务部门审批确认后入湖，并按照最小化原则严控数据访问权限。

入湖需提供信息如下：

表 2 数据入湖需求要素

信息	要求
业务系统	数据归属业务系统

数据归类	业务条线分类及数据分类
数据库连接信息	数据库类型、IP、端口
数据表信息	归属 schema、表结构说明（需包含字段中文注释）
数据特性	数据更新特性（适于定时采集还是依赖其他标识条件）、是否可增量采集、分区/增量字段、历史数据处理方式（是否可补采，时间范围，是否归档至其他库表）

7.2 技术准备

- a) 网络开通：按需梳理入湖所需网络访问关系，申请开通网络。
- b) 数据库用户：常规数据使用统一规划的入湖采集用户hxdpreader，由相关业务系统运维负责人建立用户并授权；特殊数据申请专用采集用户，系统间不共用。
- c) 文件共享：对于需要使用文件共享的场景，建立NAS专用用户hxdpuser，做好文件管理规划。

7.3 入湖实施

- a) 技术验证：各类数据入湖方案，前期进行技术验证，确认可行性、实施细节、并验证数据一致性。
 - b) 技术实施：一次性工作完成实施，日常性工作，配置平台任务。
 - c) 入湖验收：检查确认入湖内容。
- 日常监控：持续关注数据质量、源系统表结构变更，及时对日常任务进行调整完善。

附 录 A
(资料性)
数据入湖要素及验收要素

A.1 数据入湖要素

表 A.1 入湖关注要素

字段名	描述
入湖库	入湖保存库
入湖表	入湖保存表
源系统	数据来源系统
数据来源描述	非在线数据，简述数据来源
源数据库类型	来源系统数据库类型
源库	来源系统库名
源表	来源数据表名
采集频次	数据采集频次，年/月/自然日/交易日/若干小时/若干分钟
采集类型	增量或全量采集
启动时间	采集任务启动时间
依赖条件	存在何种依赖条件
历史是否补采	是否已完成历史数据补采
入湖完成时间	完成入湖时间
是否全字段	是否全字段采集
数据开始日期	数据保存开始时间
数据结束日期	离线数据截止时间
数据存储周期	数据保存周期（永久/n 天）

A.2 数据验收要素

表 A.2 入湖验收要素

验收点	描述
命名规范	入湖数据库表、入湖任务符合命名规范要求
数据完整性检查	数据记录是否完整，每日数据分布是否合理； 数据元素是否完整，字段是否完整
数据一致性检查	数据记录数及数据内容，与源系统是否一致
数据及时性检查	数据入湖时间是否符合预期

参 考 文 献

- [1] JR/T 0304.1—2024 证券期货业基础数据元规范第1部分：基础数据元
- [2] Q/HXZQ 001—2024 华西证券湖仓一体化大数据中心数据存储技术规范