2023年中国湖仓一体技术与产业研究报告

目　录

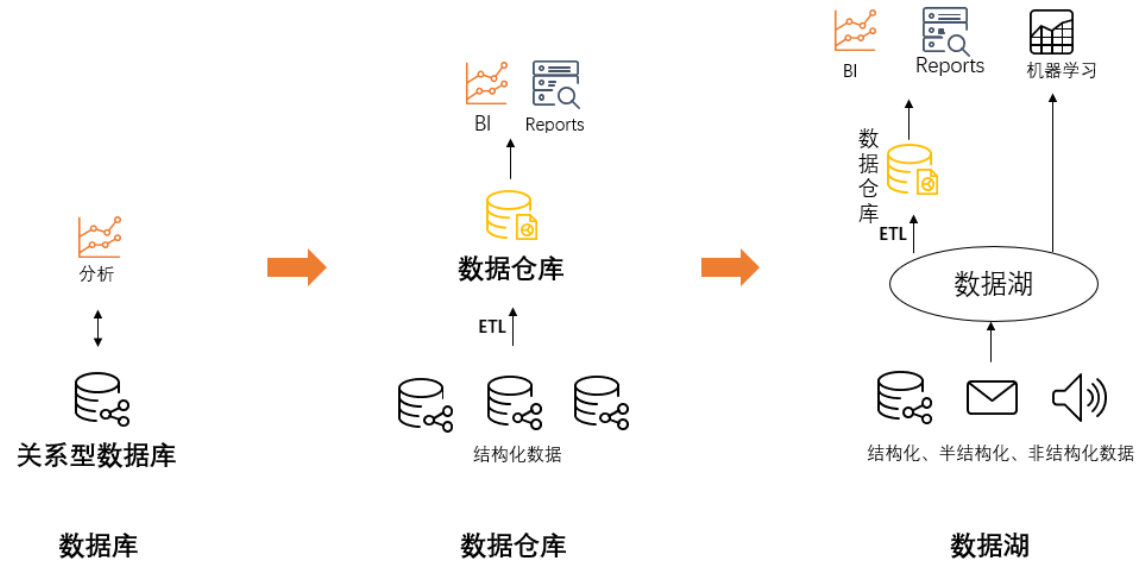
图 目 录

表 目 录

一、湖仓一体是数据平台发展的重要趋势

（一） 数据平台的发展历程

需求催生技术革新，在存储海量数据需求的推动下，数据平台架构持续演进，经过数十年的发展，主要经历了数据库、数据仓库、数据湖三个阶段。



来源：CCSA TC601

图 1 数据平台发展历程图

数据库：20 世纪 60 年代，数据库诞生，此时企业的数据量不大且数据类型比较单一。这一阶段企业对数据的使用需求主要是面向管理层从宏观层面对公司的经营状况做描述性分析，处理的数据为有限的结构化数据，支撑数据存储和计算的软件系统架构比较简单。20 世纪 70 年代，最早出现的关系型数据库已经得到了一定程度的应用。关系型数据库主要应用于联机事务处理 OLTP 场景，如银行交易等。代表产品有 Oracle、SQL Server、Mysql 等。

数据仓库：随着互联网的快速普及，门户、搜索引擎、百科等应用用户快速增长，数据量呈爆发式增长，原有的单个关系型数据库架构无法支撑庞大的数据量。20 世纪 90 年代数据仓库理论被提出。数据

仓库是为解决单个关系型数据库架构无法支撑庞大数据量的数据存储问题而诞生。数据仓库是为了对数据整合而形成的架构，核心是基于 OLTP 系统的数据源，根据联机分析处理 OLAP 场景诉求，将数据经过数仓建模形成 ODS、DWD、DWS、DM 等不同数据层，每层都需要进行清洗、加工、整合等数据开发（ETL）工作，并最终加载到关系型数据库中。数据仓库多为 MPP（Massively Parallel Processor）架构，代表产品有 Teradata、Greenplum、Clickhouse 等。

2003-2006 年，Google 的“三驾马车”：分布式文件系统 GFS、分布式计算框架 MapReduce 和数据库 Big Table，为技术界提供了一种以分布式方式组织海量数据存储与计算的新思路。受此启发开源大数据项目 Hadoop 诞生了。2008 年基于 Hadoop 自建离线数据仓库(Hive)成为数据仓库的首选方案。2010 年前后，云厂商纷纷推出云数据仓库产品，如：AWS Redshift、Google BigQuery、Snowflake、MaxCompute 等。

数据湖：随着移动互联网的飞速发展，半结构化、非结构化数据的存储、计算需求日益突出，对数据平台提出了新的要求。2010 年，数据湖概念被提出，数据湖是一种支持结构化、半结构化、非结构化等数据类型大规模存储和计算的系统架构。随着 Hadoop 技术的成熟与普及，企业开始基于 Hadoop、Spark 及其生态体系中的配套工具搭建平台处理结构化、半结构化数据，同时利用批处理引擎实现数据批处理。而以开源 Hadoop 体系为代表的开放式 HDFS 存储、开放的文件格式、开放的元数据服务以及多种引擎（Hive、Presto、Spark 等）协同工作的模式，形成了数据湖的雏形。Hudi、Delta Lake 和 Iceberg

三大开源数据湖技术的成熟，加速了数据湖产品化落地。数据湖将数据管理的流程简化为数据入湖和数据分析两个阶段。数据入湖即支持各种类型数据的统一存储。数据分析则以读取型 Schema(schema on read)形式，极大提升分析效率。代表产品有亚马逊-S3、LakeFormation，阿里云-数据湖构建 DLF、数据开发治理 Dataworks、对象存储 OSS、开源大数据平台 EMR，华为云- FusionInsight MRS 云原生数据湖、DataArts Studio 数据治理中心，腾讯云-数据湖计算服务 DLC、数据湖构建 DLF、对象存储 COS 等。

（二） 数据湖、数据仓库特性分析

数据仓库主要用于解决单个关系型数据库架构无法支撑庞大数据量的数据存储问题，很好地解决了 TB 到 PB 级别的数据处理问题，但是由于数据仓库仍以结构化数据为主，无法解决业务增长带来的半结构化、非结构化数据的存储、处理问题，且其整个建设过程需要遵循一系列规范，比如标准化的数据集成模式和存储格式、统一的数据仓库分层分域模型以及指标体系建设等，带来了数据仓库建设存储成本高、维护开发难度大、扩展能力受限制等问题。

数据湖的出现很好解决了数据仓库建设存在的一系列问题，将数据管理的流程简化为数据入湖和数据分析两个阶段。数据湖支持各种类型数据的统一存储。数据分析则以读取型(schema on read)形式，极大提升分析效率。然而数据湖对多样类型数据的支持以及灵活高效的分析方式，带来了数据治理难的问题，比如因为缺乏治理导致数据质量下降、数据不可用等，很容易退化形成数据沼泽。

总的来看，数据仓库具备规范性，可针对结构化数据进行集中式的存储和计算，但成本相对昂贵且无法处理半结构化、非结构化数据，扩展性一般、扩展成本高；数据湖具有更大的存储量，支持对于多种类型数据的高效取用，但不支持事务处理、数据质量难以保障，且缺乏一致性、隔离性。数据仓库和数据湖是两套相对独立的体系，各有优劣势，无法相互替代。

表 1 数据湖与数据仓库对比表

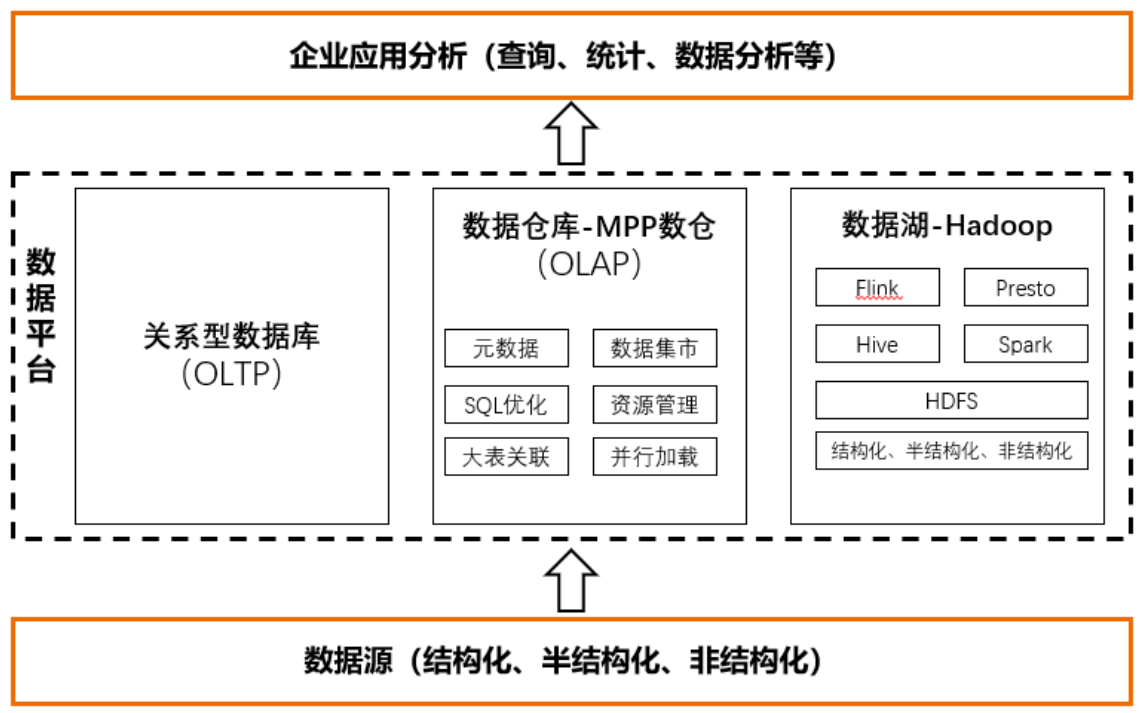
差异项	数据湖	数据仓库
数据类型	所有数据类型	历史的、结构化的数据
Schema	读取型 Schema	写入型 Schema
计算能力	支持多计算引擎用于处理、分析所有类型数据	处理结构化数据，转化为多维数据、报表，以满足后续高级报表及数据分析需求
成本	存储计算成本低，使用运维成本高	存储计算绑定、不够灵活、成本高
数据可靠性	数据质量一般，容易形成数据沼泽	高质量、高可靠性、事务隔离性好
扩展性	高扩展性	扩展性一般，扩展成本高
产品形态	一种解决方案，配合系列工具实现业务需求，灵活性更高	一般是标准化的产品
潜力	实现数据的集中式管理，能够为企业挖掘新的运营需求	存储和维护长期数据，数据可按需访问

来源：CCSA TC601

（三） 湖+仓混合业务架构存在四大痛点

为满足多种数据类型存储、多场景分析等业务诉求，企业的数据

平台采用混合部署模式，数据湖、数据仓库、关系型数据库等多种架构并存，其中数据湖和数据仓库通过 ETL 进行数据交换。数据湖和数据仓库是两套独立的体系，其中数据湖基于 Hadoop 技术生态（HDFS、Spark、Flink 等技术）来实现，主要用于支撑多源异构的数据存储，执行批处理、流处理等工作负载。数据仓库主要基于 MPP 或者关系型数据库来实现，主要支撑结构化数据在 OLAP 场景下的 BI 分析和查询需求。



来源：CCSA TC601

图 2 湖+仓混合架构图

“数据湖+数据仓库”混合架构满足了结构化、半结构化、非结构化数据高效处理需求，解决了传统数据仓库在海量数据下加载慢、数据查询效率低、难以融合多种异构数据源进行分析的问题，但也存在四大弊端：

一是数据冗余，增加存储成本。数据湖(Hadoop 技术体系)和数据仓库（MPP 技术体系）都属于分布式系统，两种技术栈都做了数据的冗余备份，同时，采用混合架构会导致部分数据既存储在 Hadoop 平

台，又存储在 MPP 平台的情况，进一步增加了数据冗余的比例，增加存储成本。

二是两个系统间额外的 ETL（抽取、转化、加载）流程导致时效性差。在数据平台实际使用过程中，数据通常先入湖，进行批处理后入仓，最后为上层应用提供查询服务，整个数据链路过长，湖入仓的过程还需进行一次 ETL，影响查询时效性。

三是数据一致性保障低，增加数据校验成本。两个系统之间通过数据迁移实现混合架构下的数据流动，在迁移过程中容易出现数据不一致问题，增加了数据一致性校验成本。

四是混合架构复杂，开发运维难度大、成本高。两种孤立技术栈混合部署使得数据架构复杂，平台开发运维难度大、成本高。

（四） 湖仓一体技术应运而生

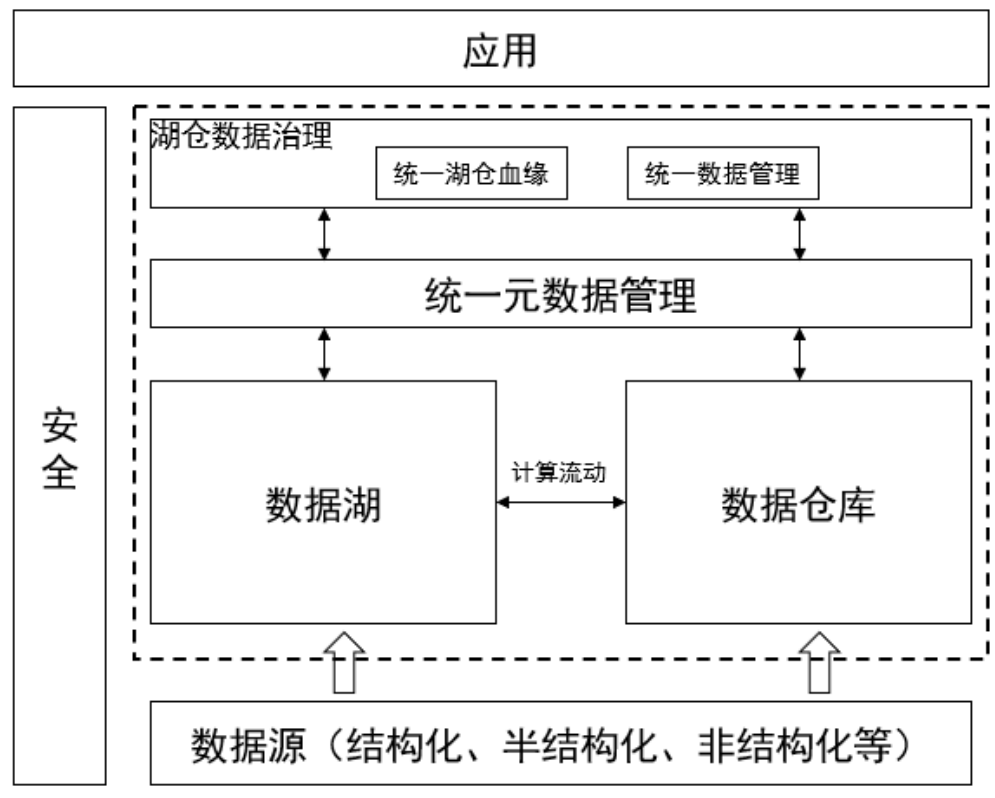
“数据湖+数据仓库”混合架构是技术向业务妥协的一个产物，并不是真正意义的湖仓一体平台。2020 年 Databricks 提出“湖仓一体”概念，随着云计算的深入应用，以容器、DevOps、微服务等为代表的云原生技术与大数据技术进一步深度融合，采用存算分离架构，同时利用云原生的资源弹性扩缩容、按需分配特点实现了资源进一步集约化，进而降低成本，同时促进了湖仓一体技术的兴起。

1. 湖仓一体概念

湖仓一体是指融合数据湖与数据仓库的优势，形成一体化、开放式数据处理平台的技术。通过湖仓一体技术，可使得数据处理平台底层支持多数据类型统一存储，实现数据在数据湖、数据仓库之间无缝

并使得上层通过统一接口进行访问查询和分析。

体架构模块图详见图 3。总的来看，湖仓一体通过引入数据仓库治理能力，既可以很好解决数据湖建设带来的数据治理难问题，也能更好挖掘数据湖中的数据价值，将高效建仓和灵活建湖两大优势融合在一起，提升了数据管理效率和灵活性。



来源：CCSA TC601

图 3 湖仓一体架构模块图

2.

为进一步规范湖仓一体数据平台技术体系，中国信通院云计算与大数据研究所依托中国通信标准化协会大数据技术标准推进委员会（CCSA TC601），联合多个电信、金融应用单位，以及阿里云、腾讯云、巨杉数据库、新华三、南大通用、甲骨文、百度云、思特奇、平安科技、云粒、科杰科技、数梦工场、滴普科技、北明数科、比智等领域内企业共同编制完成了《湖仓一体数据平台技术要求》，旨在帮助大数据产品供应商及用户方评估湖仓一体数据平台的技术能力和研发方向。本标准覆盖了湖仓一体数据平台所具备的一系列能力，总

他能力五个能力域。

湖仓数据集成	湖仓存储	湖仓计算	湖仓数据治理	湖仓其他能力
数据源管理	存算分离	存储生态支持	统一元数据管理	异地容灾
湖仓数据转换能力	存储分级	认证授权	统一数据管理	
入湖仓能力	数据湖格式	统一开发平台	统一湖仓血缘	
	存储加速	弹性能力	数据评估能力	
	存储加密	多场景融合分析	数据标准及数据质量	
		统一资源管理	动态数据加密	
		多计算模式支持	数据建模能力	

来源：CCSA TC601

图 4《湖仓一体数据平台技术要求》标准总体框架

2.1湖仓数据集成能力

便利的数据入湖、入仓是湖仓一体纳管数据能力的开始。湖仓数据集成能力包括（1）统一外部关系型数据库、NoSQL 数据库、分布式文件系统等数据源的管理。(2) 数仓可对数据湖数据对象转换为数仓的数据管理对象进行数据和权限管理（升仓），同时支持数仓内价值密度低的数据进行入湖操作的湖仓数据转换能力。(3) 具备实时与批量数据入湖、入仓能力，以及入湖任务配置与管理的入湖仓能力。

2.2湖仓存储能力

湖仓存储需兼容数据格式，保障数据自由入湖仓的安全和质量。湖仓存储能力包括（1）具备数据存储和计算资源独立部署，以及动态扩缩容存储、计算资源的存算分离能力。(2) 湖仓数据冷、热分级存储的存储分级能力。(3) 支持 Hudi、Iceberg、Deltalake 等数据湖格

,支持模式（schema)在线调整。（4）数据缓存加速能力，支持配置多种缓存策略的存储加速能力。（5）湖仓数据加密存储的存储加密能力。

2.3湖仓计算能力

湖仓一体架构涉及异构数据平台对数据的处理，与传统ELT/ETL 形式不同的是数据无需移动。湖仓计算能力包括（1）存储生态能力，涵盖数仓引擎可以对数据湖数据进行读写，数据湖引擎同样可对数仓数据进行读写。（2）统一的认证、授权体系。（3）统一开发平台进行湖仓数据开发利用、作业调度、任务运维监控。（4）计算资源弹性扩缩容，且能够对弹性资源的使用情况进行监控。（5）对湖仓数据可进行科学计算、向量计算、机器学习等多场景融合分析。(6)对湖仓存储资源、计算资源进行统一管理、分配、使用以及监控。(7)支持批处理、实时计算、OLAP 分析等多种计算模式。

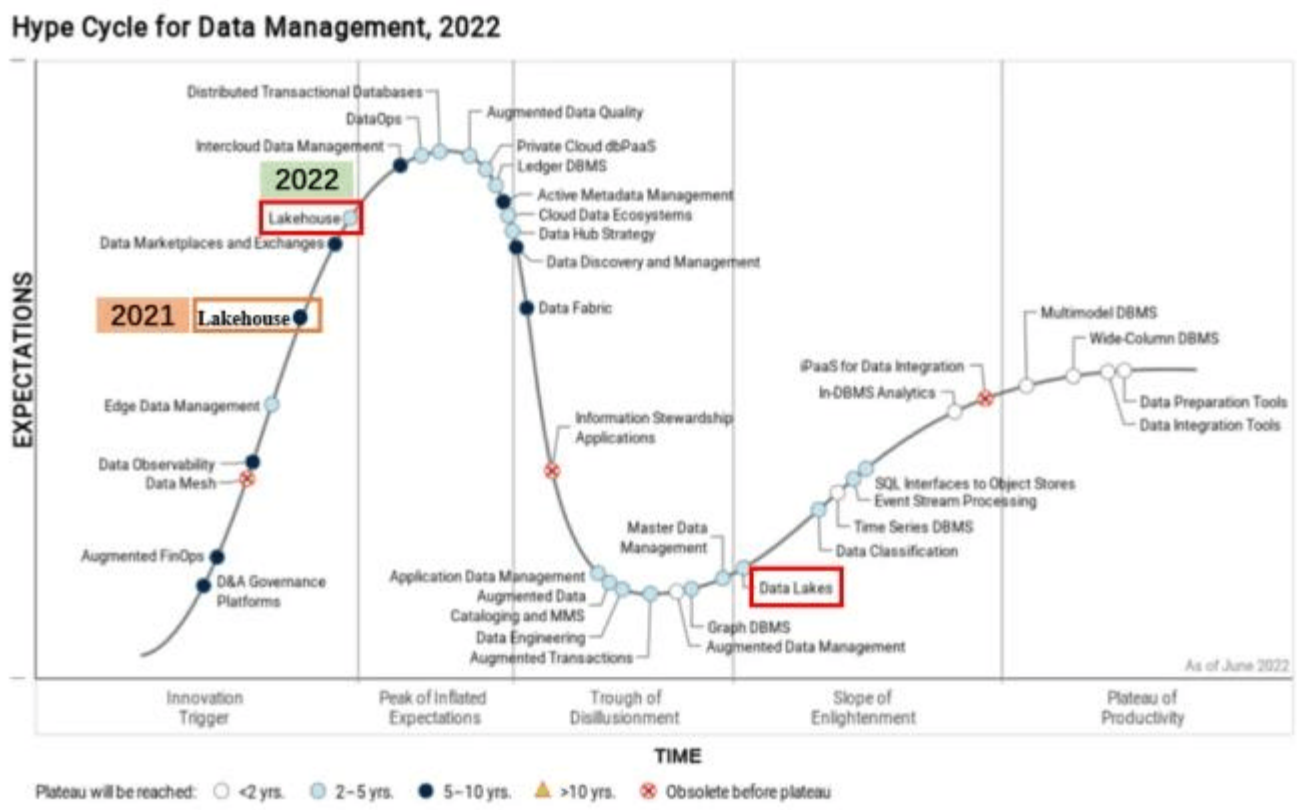
2.4湖仓数据治理能力

统一数据治理能够替客户屏蔽底层异构数据平台的复杂性，给客户带来更好的体验。湖仓数据治理能力包括（1）元数据自动发现、自动识别、自动采集、元数据存储等统一元数据管理能力。（2）对湖仓内数据有统一的数据权限管理能力。(3)对数据的访问频次、时间、数据量等维度可进行评估的数据评估能力。(4)对湖仓内的数据流转、生命周期有清晰描述的统一湖仓血缘能力。（5）支持数据质量的规则设置、校验以及质量管理。（6）可在湖仓异构访问过程中对敏感数据加密。（7）可提供统一数据建模能力，包含逻辑模型、物理模型，并

2.5湖仓其他能力

本标准梳理了湖仓一体必备且专有的技术要求能力，除去存储、计算、集成、治理外的其他能力，主要包括异地容灾能力。

自 2021 年“湖仓一体”首次写入 Gartner 数据管理领域成熟度模型报告以来，湖仓一体技术备受关注。从 Gartner 发布的《Gartner 数据管理成熟度曲线》(2022年)可以看出，数据湖技术日趋成熟，湖仓一体技术成熟期相比 2021 年缩短，期望值升高。同时各大云厂商纷纷推出湖仓一体产品，如 AWS 智能湖仓、Databricks- Lakehouse Platform、阿里云- MaxCompute 湖仓一体、华为云- FusionInsight MRS、腾讯云-云原生智能数据湖。



来源：Gartner

图 5《Gartner 数据管理成熟度曲线》2022 年

企业需求的驱动下，数据湖与数据仓库在原本的范式之上向其限

实现路径。湖上建仓和仓外挂湖虽然出发点不同，但最终湖仓一体的目标一致。如表 2 所示，展现了两种路径在优劣势、实现方向、亟需解决问题等维度的对比。本章节将详细介绍两种实现路径。

表 2 两种实现路径对比表

实现路径	优势	劣势	需解决的问题	实现方向
湖上建仓（Hadoop 体系）	支持海量数据离线批处理	不支持高并发数据集市、即席查询、事务一致性等	1.统一元数据管理 2.ACID 3.查询性能提升 4.存储兼性问题 5.存算分离 6.弹性伸缩	1.提升查询引擎、存储引擎能力
仓外挂湖（MPP 体系）	事务一致性，结构化数据 OLAP 分析	不支持非结构化/半结构化数据存储、机器学习等	1.统一元数据管理 2.存储开放性 3.扩展查询引擎 4.存算分离 5.弹性伸缩	1.计算引擎不变，只扩存储能力。 2.查询引擎扩展，提升查询引擎效率

来源：CCSA TC601

湖上建仓

湖上建仓是指基于云存储或第三方对象存储的云数据湖架构，或者基于开源 Hadoop 生态体系并以 DeltaLake、Hudi、Iceberg 三大开源数据湖作为数据存储中间层实现多源异构数据的统一存储，以统一调用接口方式调用计算引擎，最终实现上下结构的湖仓一体架构。代表产品有：华为云-FusionInsight MRS、AWS-智能湖仓、Databricks - Delta Lake 等。

基于开源 Hadoop 生态体系，擅长海量数据离线批处理，在高并

现途径中，实现方向为提升查询引擎、存储引擎能力。

总的来看“湖上建仓”路径本质是在湖的基础上增加仓的能力，需解决以下六大技术难点：

元数据的统一最为核心，是确保湖仓一体在架构和应用层面达到统一的关键。湖上建仓路径通过增加元数据管理组件实现元数据的统一管理，目前大都只实现了元数据的采集和统一存储。

二是事务支持。湖上建仓通过集成 Hudi、Iceberg、Delta Lake 三大开源数据湖表格式进行优化，支持数据更新，实现支持事务的存储层。

三是提高查询性能。湖上建仓路径在引擎加速和存储优化方面，通过引入如缓存加速、谓词下推、元数据相关语义优化、C++重写引擎等能力来解决原有计算、存储引擎的性能瓶颈问题。

四是存储兼容性。湖上建仓路径中的存储介质由原有的以 HDFS 为主，扩展到支持云对象存储等多种介质存储。

五是存算分离。传统的 Hadoop 体系不具备云原生能力，是存储和计算部署在同一物理集群来应对网速不足、数据在各节点间交换时间长的问题。湖上建仓则是将 HDFS+对象存储独立部署，实现存算分离。

六是弹性伸缩。基于 K8S、Docker 等容器化技术对 Hadoop 体系组件、服务进行容器化改造。目前大部分产品有实现计算层、存储层

弹性伸缩，少量产品实现了根据业务负载自动弹性伸缩计算资源。

（二） 仓外挂湖

仓外挂湖是指以 MPP 数据库为基础，使用可插拔架构，通过开放接口对接外部存储实现统一存储，在存储底层共享一份数据，计算、存储完全分离，实现从强管理到兼容开放存储和多引擎。代表产品：Snowflake、AWS Redshift、阿里云 MaxCompute/Hologres 湖仓一体。

MPP 数据库技术体系，从关系型数据库演进而来，对事务一致性、联机分析处理性能都有较好的支撑，但在分析场景方面存在较大的局限性，主要以结构化数据分析为主，无法支撑半/非结构化数据存储、实时计算、机器学习等场景。所以实现途径中，实现方向为增加存储能力，提升查询引擎效率。

总的来看，“仓外挂湖”路径本质是在仓的基础上增加湖的多类型存储等能力，需解决以下五大技术难点：

一是统一元数据管理。打通不同数据系统，具备数据共享和跨库分析的能力，并支持互联互通、计算下推、协同计算，实现数据多平台之间透明流动。仓外挂湖路径目前主要是将对接外部存储如 Hadoop、对象存储等的元数据进行采集，统一存储、管理。

二是存储开放性。仓外挂湖路径的存储开放性主要表现在：存储介质兼容方面，将非数仓自身存储如 Hadoop、云对象存储等的数据纳入管理；数据格式方面，采用开放、标准化的数据格式，既包含 Hudi、Iceberg、Delta Lake 等开放格式，也包括 Parquet、ORC、CSV 等存储