大数据技术标准推进委员会
Big Data Technology and Standard Committee

湖仓一体技术与产业研究报告

(2023 年)

CCSA TC601 大数据技术标准推进委员会

2023年6月

版权声明

编制说明

本报告的撰写得到了大数据领域多家企业与专家的支持和帮助，主要参与单位与人员如下。

参编单位：大数据技术标准推进委员会、中国移动通信集团山东有限公司、威海市商业银行、阿里云计算有限公司、腾讯云计算（北京）有限责任公司、华为云计算有限公司、北京镜舟科技有限公司、北京飞轮数据科技有限公司、科大讯飞股份有限公司、中兴通讯股份有限公司、天津南大通用数据技术股份有限公司、杭州数梦工场科技有限公司、杭州比智科技有限公司、北京数势云创科技有限公司、浙江数新网络有限公司、北京百度网讯科技有限公司、北京滴普科技有限公司、北京科杰科技有限公司、北京偶数科技有限公司。

参编人员：魏凯、姜春宇、闫树、马鹏玮、田稼丰、刘彦美、朱祥磊、高鹏、魏冲、刘一鸣、孔亮、林楠、王宁、崔潇扬、杨勇强、汪定新、王涵毅、白雪、赵峰、汤雅琴、伍攀、陈关良、赵青柏、谢辉、高经郡、张立群。

前　言

数据平台是能够为企业提供数据分析能力、支撑上层数据应用、助力企业数字化转型的底层基础设施，它包含数据存储、数据计算分析等能力的一套基础设施，通过汇聚各方数据，提供“采-存-算-管-用”全生命周期的软件支撑。经过数十年的发展，数据平台架构持续演进，主要经历了数据库、数据仓库、数据湖三个阶段。

如今，数据仓库和数据湖是数据平台最广泛的两种架构：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，但无法处理半结构化与非结构化数据，且其扩展能力有一定局限性；数据湖具有更好的扩展能力，能够灵活支持对于多种类型数据的高效取用，但不支持事务处理，缺乏一致性、隔离性，数据质量难以保障。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。

为满足多种数据类型存储、多场景分析等业务诉求，企业采用数据湖+数据仓库混合架构。“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，但也存在混合架构复杂，开发运维难度大、成本高，数据处理链路长时效低等问题。

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、数据仓库之间无缝调度和管理，并使得上层通过统一接口进行访问查询和分析。

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，随着企业数字化转型的不断深入，“湖仓一体”作为新型的技术受到了前所未有的关注，越来越多的企业视“湖仓一体”为数字化转型的重要基础设施。湖仓一体平台的建设解决了流批一体面临的原子事务、一致性更新以及元数据性能瓶颈等问题，使得湖仓一体平台的构建既能满足短期业务发展的需要，又能支撑长期的数据应用诉求。

为给社会各界深入了解湖仓一体技术与产业提供有价值的参考。本报告聚焦于湖仓一体技术，详细梳理了数据平台发展历程、湖仓一体实践路径，研究分析了湖仓一体产业现状，并对湖仓一体未来发展进行了展望与研判。由于时间仓促，水平所限，错误和不足之处在所难免，欢迎各位读者批评指正，意见建议请发送至 liuyanmei@caict.ac.cn。

目 录

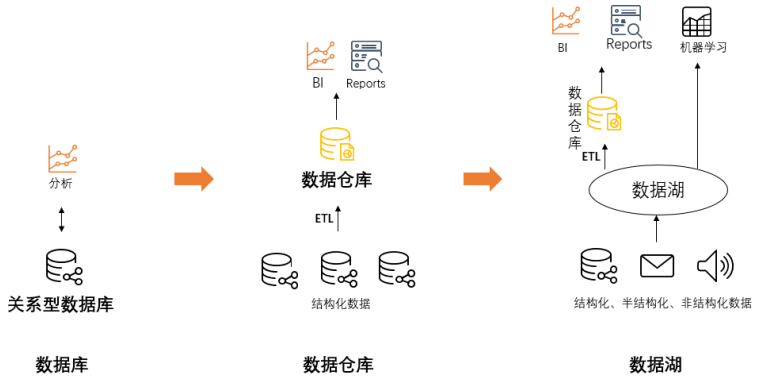
图 目 录

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

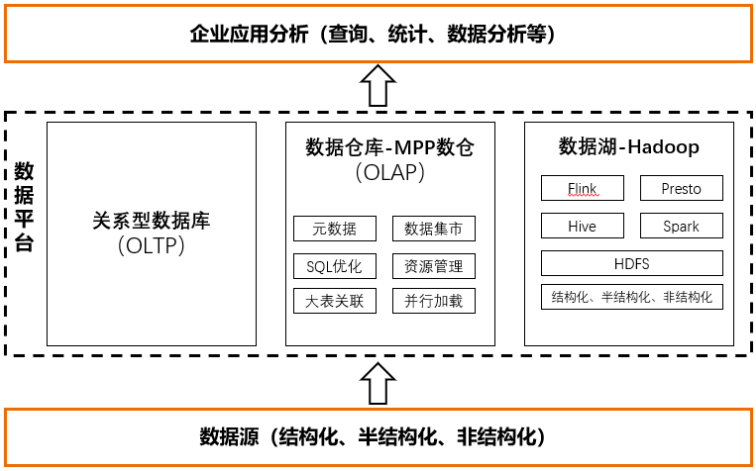
差异项	数据湖	数据仓库
数据类型	所有数据类型	历史的、结构化的数据
Schema	读取型 Schema	写入型 Schema
计算能力	支持多计算引擎用于处理、分析所有类型数据	处理结构化数据，转化为多维数据、报表，以满足后续高级报表及数据分析需求
成本	存储计算成本低，使用运维成本高	存储计算绑定、不够灵活、成本高
数据可靠性	数据质量一般，容易形成数据沼泽	高质量、高可靠性、事务隔离性好
扩展性	高扩展性	扩展性一般，扩展成本高
产品形态	一种解决方案，配合系列工具实现业务需求，灵活性更高	一般是标准化的产品
潜力	实现数据的集中式管理，能够为企业挖掘新的运营需求	存储和维护长期数据，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

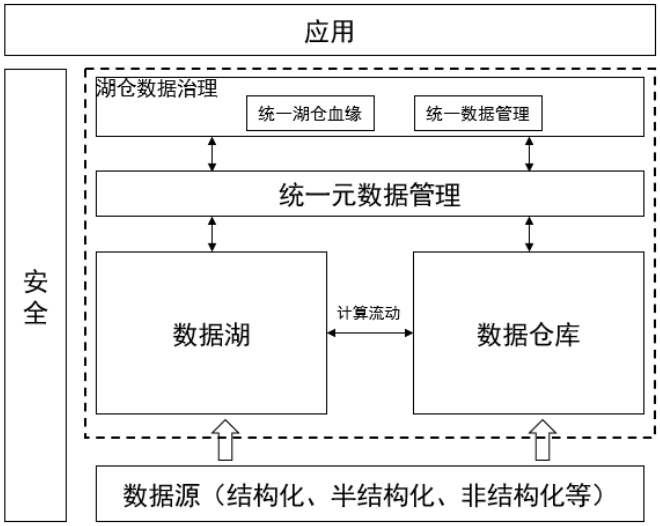
（四）湖仓一体技术应运而生

“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

调度和管理，并使得上层通过统一接口进行访问查询和分析。湖仓一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. 湖仓一体基本能力

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

体分为**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**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

式，且实现事务支持处理能力,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

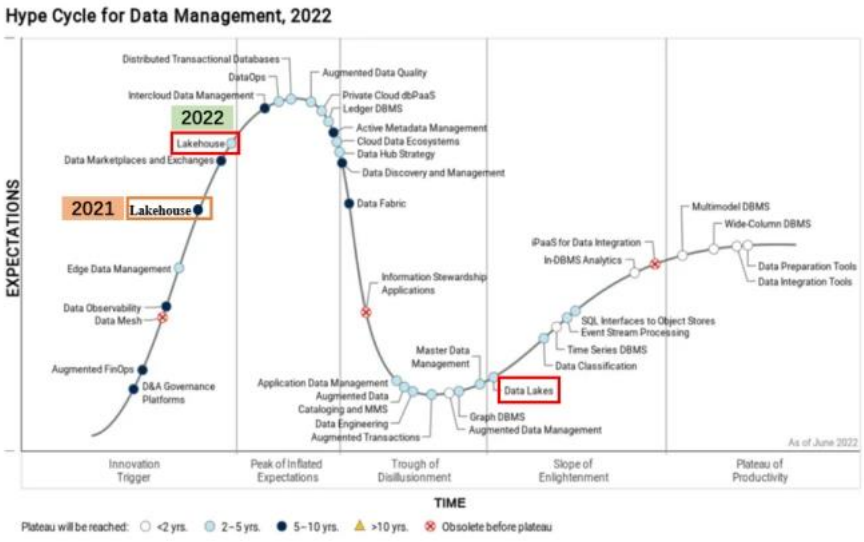
统一数据治理能够替客户屏蔽底层异构数据平台的复杂性，给客户带来更好的体验。湖仓数据治理能力包括（1）元数据自动发现、自动识别、自动采集、元数据存储等统一元数据管理能力。（2）对湖仓内数据有统一的数据权限管理能力。（3）对数据的访问频次、时间、数据量等维度可进行评估的数据评估能力。（4）对湖仓内的数据流转、生命周期有清晰描述的统一湖仓血缘能力。（5）支持数据质量的规则设置、校验以及质量管理。（6）可在湖仓异构访问过程中对敏感数据加密。（7）可提供统一数据建模能力，包含逻辑模型、物理模型，并

提供数据模型的生命周期管理。

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

二、湖仓一体实践路径

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

制范围扩展，逐渐形成了“湖上建仓”与“仓外挂湖”两种湖仓一体实现路径。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

（一） 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

发数据集市、即席查询、事务一致性等方面存在先天的不足。所以实现途径中，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:

一是统一元数据管理。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

格式的支持。

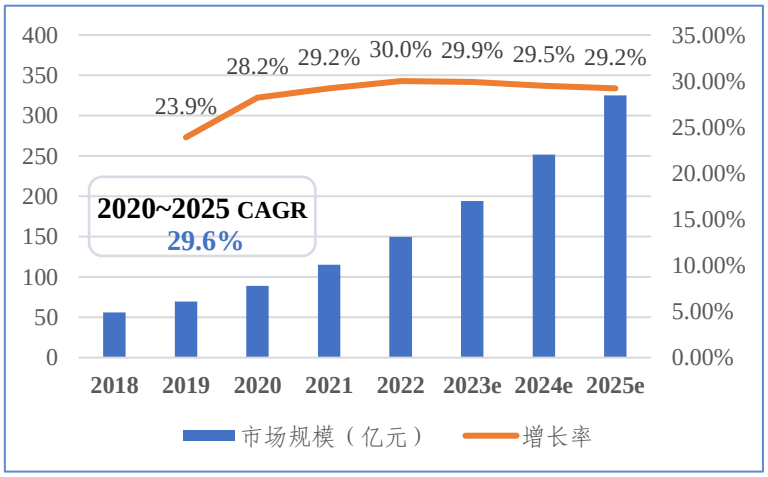
三是扩展查询引擎。仓外挂湖路径保留原 MPP 计算引擎计算能力的基础之上，主要是增加批处理和实时数据处理的能力。其中批处理方面是融合更轻量级、高效率的计算能力，而实时处理方面则是通过微批以及增量计算的方式，增强流的计算能力。

四是存算分离。仓外挂湖需进行存算分离架构改造，而传统的 MPP 存算耦合架构，不具备云原生能力。目前，仓外挂湖路径主要基于存算分离架构改造后的云原生 MPP 数据库实现。

五是弹性伸缩。基于 K8S、Docker 等容器化技术对 MPP 体系的组件、服务进行容器化改造。目前该路径有实现计算层、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。

三、湖仓一体产业及应用现状

随着企业数字化转型驱动市场需求的不断增加，同时开源技术的发展降低了企业加入大数据领域的门槛，加之数据量的规模化增长和应用场景的越发丰富，数据平台需求不断扩大，数据平台软件市场稳步增长。据 CCSA TC601 测算，未来三年我国数据平台软件市场以**29.6%**的复合增长率快速发展，2025 年我国数据平台软件市场规模将超 **300 亿元**。



来源：CCSA TC601

图 6 我国数据平台软件市场规模

（一）湖仓一体主要厂商和代表产品

自 2020 年湖仓一体概念被提出，阿里云、华为云、亚马逊云等云厂商纷纷提出自己的湖仓一体架构理念，于 2021 年陆续发布湖仓一体产品。导致这一现象的原因：一方面，云厂商先发优势，云计算的弹性算力、数据聚合等能力与湖仓一体的一体化思路相符合。另一方面，在布局实践上云厂商率先基于云原生理念在对象存储、多模计算、统一管理等湖仓一体核心技术上进行了能力整合，服务自身业务诉求。表 3 整理了目前国内外湖仓一体主要厂商、代表产品。

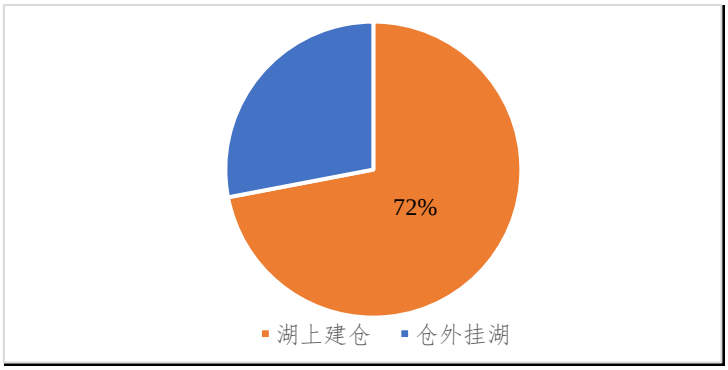
表 3 湖仓一体主要厂商和代表产品

厂商	湖仓一体
阿里云	MaxCompute/Hologres 湖仓一体
HUAWEI	FusionInsight MRS 云原生数据湖
腾讯云	云原生智能数据湖
移动云	云原生大数据分析 Lakehouse
TRANSWARP 星环科技	TDH

火山引擎	LAS
keer1科杰科技	KeenData Lakehouse
H3C	H3C 绿洲融合集成&数据运营平台
GBASE 南大通用	GCDW
inspur 浪潮科技	行业数字平台
BONC 东方国信	企业级数据湖（BELAKE）
滴普科技 DEEPEXI	FastData
海康威视 HIKVISION	海康威视大数据基础平台
SELECTDB	SelectDB
OUSHU 偶数	Oushu Data Cloud
aws	AWS 智能湖仓
databricks	Lakehouse Platform
snowflake	Snowflake

来源：CCSA TC601

根据 CCSA TC601 统计分析，目前国内七成以上厂商基于“湖上建仓”实现路径，如华为云-FusionInsight MRS 云原生数据湖、腾讯云-云原生智能数据湖、移动云-云原生大数据分析 Lakehouse 等，近三成厂商基于“仓外挂湖”实现路径，如阿里云-MaxCompute/Hologres 湖仓一体等。

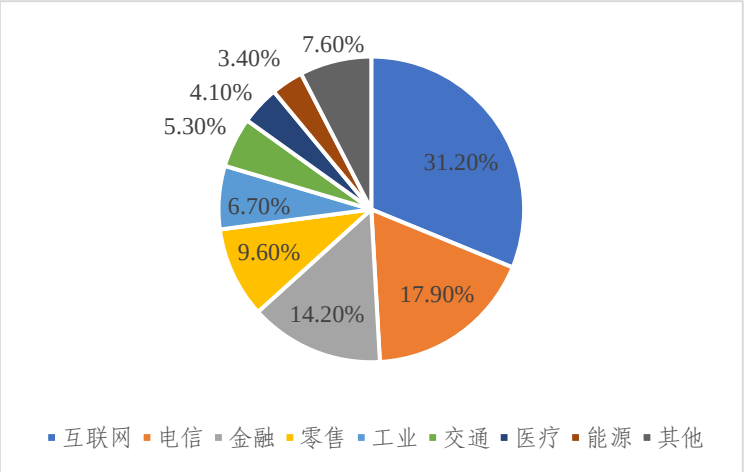


来源：CCSA TC601

图 7 实践路径统计图

（二）湖仓一体在互联网、电信、金融等信息化程度高的领域应用程度高

互联网、电信、金融行业是国内数字化程度较高的领域，数据管理体系相对完善，更加重视数据的使用、挖掘、分析、存储等能力。从图 7 中可以看出，湖仓一体的应用领域主要集中在互联网、电信、金融等行业，占比超过六成。



来源：CCSA TC601

图 8 2022 年湖仓一体市场行业统计图

随着中共中央国务院印发《数字中国建设整体布局规划》，指出建设数字中国是数字时代推进中国式现代化的重要引擎，未来湖仓一体平台将在政府、工业、交通等数据价值尚未完全释放的行业加大应用渗透率。

表 4 各行业需求现状表

行业	需求及现状
互联网	互联网企业不断产生各种新应用，数据来源多样，导致需要从海量数据中分析得到有价值的信息数据，进而辅助企业决策。湖仓一体平台可加快挖掘价值数据的速度，减少数据存

	储成本，支撑业务快速迭代发展。
电信	电信行业拥有庞大的个人位置数据，有精准营销、信用评估等应用诉求。目前采用的传统数据架构，存在数据质量不高、实时性不够、灵活性不足、存储应用相互制约等问题。湖仓一体平台实现了可规模化、低成本存储，同时可根据应用按需建模，推动了行业的垂直应用。
金融	金融行业数据资产化更为成熟，信息化建设起步早、资金投入巨大，数据标准化程度高，且技术实力强。目前依赖的传统数据基础设施无法处理金融机构目前收集的各种海量数据，而且个性化客户参与和降低风险的需求日益增长。湖仓一体平台实现了统一存储、大规模分析等能力，满足创新业务需求，提升用户体验。
零售	零售行业中个性化买家细分和基于客户行为的实时体验需求不断增长，随着线上线下各种零售渠道的涌现，线下门店、自有商城、电商平台、社交软件平台等渠道也带来大量碎片化的数据。湖仓一体平台打通企业内外部数据，实时更新“进-销-存”数据，进而实现智能化分析。
政府	在政策的驱动下，政府部门的信息化建设投入不断加大，基础设施建设已经趋于成熟。目前以智慧城市/政务为中心的信息化建设正在加速推进，需解决政务服务业务创新速度落后于社会需求的问题，推动数据与业务的融合，提升服务型政府供给侧能力。
工业	工业数据价值高，标准与治理痛点突出，处于数字化转型的关键时期，由于工业生产制造流程复杂且专业性强，而且目前数据基础设施建设薄弱，湖仓一体平台可帮助工业进行数据统一治理，未来在工业领域仍有较大的发展空间。
交通	交通行业处于数字化转型的起步阶段，其业务特性决定了具有较高的智能化应用潜力。目前部分业务环境（如智能交通、智慧机场等）应用了数据平台，缓解交通拥堵、改善城

	市交通状况，提升交通智慧化水平和运营效率。
医疗	医疗数据大多是非结构化数据，医生、医院、患者等各方面都极为重视医疗数据的安全存储能力，传统数据库已无法满足医疗行业临床业务的数据管理及存储需求。目前也在不断作数字化转型尝试，比如用于处方、诊断等医学信息的自动理解与提取，帮助医疗人员进行信息整合。
能源	能源行业信息化建设起步较晚，前期主要支撑各业务系统运行，随着企业对数据管理和应用的重视，其数字化进程也在不断加快。

来源：CCSA TC601

四、结论与展望

湖仓一体行业正处在发展初期，总的来看湖仓一体并不是一个纯技术攻关工作，而是技术逐步融合、整合的过程，其本质是异构数据平台走向一体化的过渡阶段。

湖仓一体的核心是实现数据湖和数据仓库中的数据、元数据的无缝打通，并可自由流动。数据湖中的“新鲜”数据可以流转到数据仓库中，甚至可以直接被数据仓库使用，而数据仓库中的“不新鲜”数据，也可以流转到数据湖中，低成本长久保存，供未来的数据挖掘使用。目前，业界在湖仓一体技术的研究主要集中在统一元数据管理、统一存储等方面，仍需持续深耕。

随着数字经济时代数据的价值被进一步重视和挖掘，各行业对新一代数据平台的需求不断扩大，湖仓一体技术欣欣向荣，具有非常广阔的发展空间。同时随着大数据、人工智能与云计算的边界越来越模糊，三者不断相互影响与融合，未来，湖仓一体呈现以下三点趋势：

一是进一步简化数据架构实现一体化。统一的数据底座可以屏蔽底层部署的复杂性，为应用层带来更一致的体验，无论是经营型还是创新型应用都能获得更高效的支持，即可一站式满足企业实时分析、交互查询、智能探索等高价值数据洞察诉求。同时为数据工程师、数据科学家、数据分析师等不同角色提供低门槛自助分析能力，使其拥有更好的数据使用体验。

二是利用云原生概念实现湖仓一体无服务器化部署。Serverless 无服务化是指湖仓一体架构中的数据存储、数据查询引擎、数据处理等均支持无服务器部署，允许用户在不构建不运维一个复杂基础设施的情况下可进行开发、运行和管理。Serverless 部署给用户带来更易用的使用体验，帮助用户更专注于业务本身，而非关心技术逻辑，此外 Serverless 部署还可提供按需计费，进而实现更高效的资源利用。

三是 AI 助力湖仓一体资源调度更顺畅。随着 AI 技术广泛应用，不仅让湖仓一体的运维、部署更加智能，还可以使得资源调度更加顺畅，从而打通数据和业务智能化之间的阻隔，实现价值闭环。智能化能力重塑了湖仓一体架构中的数据供给和管理方式，可实现敏捷数据洞察和高效一致的数据协作，能够以更低的成本、更迅速地做出可信业务决策，实现 10 倍以上的数据化运营效率的提升。

附录：典型案例

（一） 山东移动：湖仓一体大数据平台建设实践

1. 案例背景

经营分析系统和大数据专题分析平台的建设将围绕经营决策工作提供更加全面、深入、高效的数据展开,在这一背景下运营商企业不断推进 B 域、O 域、M 域的数据融合，传统经分系统和大数据平台也随之需要承载更大的数据量和业务量。

2. 拟解决的痛点、难点

山东移动湖仓一体实践拟解决痛点聚焦于以下三个方面:

1）**业务数据未完全实现实时融通问题**，不同分析诉求，不同数据分散在不同的存储中，访问接口不一致;

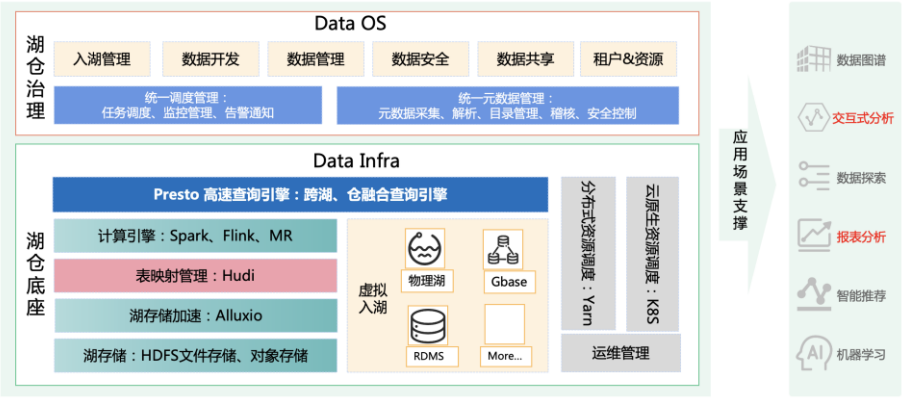
2）**数据访问慢，批量入仓压力大问题**，传统数仓数据更新采用左右关联和 Insert Overwrite 分多个步骤实现，流程长，效率低;

3）**应用实时性需求难以满足**，实时业务流程长，目前实时需求支持还是准实时的微批量数据处理，数据处理流程长，实时性不足。

3. 解决方案

山东移动湖仓一体采用“湖上建仓”路径，具体解决方案是以经分大数据平台为基础提供统一元数据管理，实时/非实时统一入湖、跨域数据统一访问、异构数据源统一计算等能力的大逻辑数据湖架构，引入亚信的 Data Infra 产品构建湖仓底座能力，Data OS 产品形成

湖仓库治理能力。



在现有集成能力基础上，引入 Hudi 组件，构建数据实时入湖能力，实现数据增量入湖，减少计算资源消耗，提高数据更新效率和业务决策速度。通过 Hudi 提供事务操作和快照的特性，提供跨层的增量更新能力，较原来的全量扫描表，在执行效率想有大幅度的提高，并在资源使用量方面也有明显的降低。基于 Alluxio 构建存算分离的架构，实现存储和计算可以独立的扩展和伸缩，保证湖仓一体的整体系统能够支持更多的用户并发和更大的数据量，同时最大程度地利用资源，从而实现对大规模数据进行查询和高效分析。通过融合元数据管理，提供了跨湖仓和仓库的元数据统一管理，在安全等级的限制下，通过同步 Hudi 的元数据到 HMS，形成统一的共享元数据中心，为组织内部的用户提供湖仓一体下的统一开发和建模。

4. 价值与效果

基于 Hudi+Presto+Hadoop+Gbase 的湖仓一体大数据平台架构，能够有效降低硬件成本，成本为之前的十分之一。在大规模海量分析场景下，性能提升 **10-20 倍**。同时存算一体的架构解决了多种架构混合使用的数据冗余问题，一份数据，实时共享，节省了大量存储成本

和人力维护成本。

湖仓一体解决了传统大数据平台面临的资源扩展灵活性低、扩容需要重分布数据，成本高时间长、数据冗余大、混合负载场景存在资源争抢等问题，通过引入基于存算分离的湖仓一体架构，实现融合 OLTP、OLAP、Hadoop 等多种数据引擎，实现多类数据引擎间的数据共享和流通互访，具备数据统一管理能力，在数据集市业务上线后，业务性能提升 **2 倍**，存储降低 **4 倍**，取得了显著效果。

（二） 威海银行：传统数据仓库到湖仓一体建设实践

1. 案例背景

威海市商业银行于 2012 年开始开展数据仓库建设（基于 DB2 数据仓库），按需实现数据集中接入和应用系统数据供给，支撑全行共性数据加工和报表统计分析及查询。但是伴随行内信息化进程加快，数据孤岛、开发周期较长、数据冗余、数据服务支撑能力弱、数据架构扩展性差和数据集群算力低等不足也逐步显现。

2. 拟解决的痛点、难点

1）提升开发运维效率

采用先进大数据及分布式数据库技术，构建适用于行内的数据架构及企业级数据平台。定制一套简单、快捷的开发平台，梳理适合开发、运维工作流程，提升开发运维工作效率。

2）提高系统计算性能

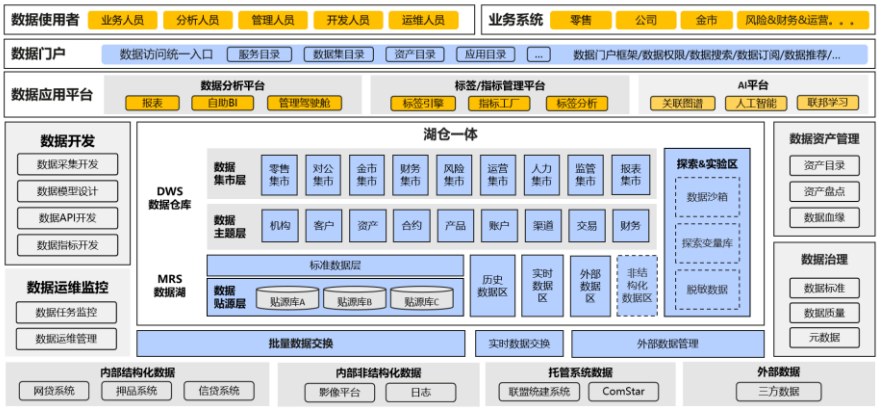
具备海量数据存储及分析处理能力，支持横向扩展，合理配置作

业并发数量，充分利用系统资源，满足日终批量处理时间要求。

3）提升数据服务能力

通过提升数据服务，采用新的数据服务模式，利用数据发布、订阅和数据 API 等方式，提升数据服务能力；整合内部数据与外部数据、流式数据与批量数据，构建企业级数据模型，全面支撑行内经营管理数据需求；通过数字化场景工作坊挖掘以客户为中心的数字应用场景。

3. 解决方案



威海银行采用基于 MRS 和 DWS 的湖仓一体方案逐步替换基于 DB2 的传统数据仓库，并最终实现湖仓一体对行内数据应用场景的全面支撑。威海银行于 2022 年 3 月启动湖仓一体项目建设，项目建设主要分为 3 个阶段：

第一阶段，夯实基础。明确定位，形成全行数据体系，建立全行统一数据架构，试点支撑数字化转型项目数据服务需求。截至 2022 年 5 月，已完成数据湖及数据仓库集群搭建和第一期数据入湖工作，支撑关联交易、贷后管理等业务系统用数需求。

第二阶段，业务赋能。形成全行数据服务体系，构建业务主题集

市，全面提升全行用数能力和水平。截至 2023 年 5 月，已构建企业级基础主题模型 200 余项，全面落实数据治理工作提出的数据标准，保障数据有序、高效、保质、安全使用；建设零售、对公、金市、风险、监管等 9 大业务集市，支撑智慧营销、财务盈利性分析和监管报送等业务应用，同时满足业务自助分析场景。

第三阶段，引领创新。持续演进湖仓一体建设，构建全面的数据智能实时服务，加深业务发展和数据服务的融合。主要开展数据应用与服务的持续优化，深度融合数据服务与业务流程、优化数据服务框架，探索数据智能应用场景，构建开放式数据服务体系等工作。

4. 价值与效果

通过湖仓一体建设实践，带来成效如下:

业务服务方面。全面支撑智慧营销、智慧运营、风险防控、监管报送等应用场景，支持 22 个数字化转型项目数据服务，智慧营销累计获客目标完成 187.5%，报表自动化率达到 88%。

能力建设方面。构建科学合理的数据架构，全面提升数据接入能力、数据整合能力和数据加工效率，在日终作业数量增加 5 倍的基础上，日终批量加工效率提升 200%；通过 BI 自助报表工具引入和自助用数培训宣贯，营造全行自助用数氛围。

平台运行方面。通过数据中台湖仓一体的建设，集群算力提升 3 倍，资源利用率提升 30%，有效支撑海量数据加工分析、模型预测等场景。

（三） 阿里云：国内某互联网金融客户湖仓一体建设实践

1. 案例背景

客户是一家互联网金融公司，2015 年成立，曾获“2019 年度竞争力金融科技创新实力公司”奖。从国外某厂商迁移到阿里云后，持续进行平台的迭代升级，建设和改造数据湖架构，构建围绕数仓的数据中台。

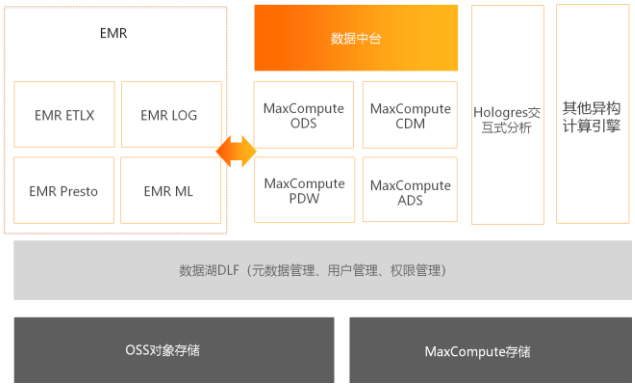
客户对数仓基础功能、安全、Serverless 云原生能力都有较高要求，并且基于数仓有 OLAP 分析服务的需求。客户原有 Hadoop 技术栈的多套业务系统，也有基于 OSS 存储，上面多套开源引擎共享数据的需求。

2. 拟解决的痛点、难点

1）用户的数据湖不能满足数仓的多租户、安全隔离、Serverless、免运维等要求。例如 Atlas 只能做表示层功能无法在引擎层统一元数据；Ranger 的安全管控能力无法让多个部门共享一套 Hadoop 平台实现多租安全隔离，并通过权限快速共享；Presto 联邦分析引擎可以交互式查询但不能做可靠的联邦数据处理。且互联网金融客户关注业务，不愿投入太多成本修改补充开源短板。

2）客户认为湖的能力不满足数仓要求，于是引入云原生数仓 MaxCompute 构建数据中台，MC 与原有开源体系异构，如果不能融合，也将成为湖上的一个组件，可能带来存储冗余、元数据不统一、权限不统一、湖仓计算不能自由流动等问题。

3. 解决方案



MC 提出湖仓一体联邦方案，经过以下步骤，将湖仓进行打通。

1）将云网络和客户部署 Hadoop 的 VPC 网络进行了打通，并提供高速缓存和网络链接对象，可以快速建立、维护、共享网络链接。

2）创建 OSS 和 VPC 中的 Hadoop 实例的外部服务对象，将数据湖实例自动映射为仓中的逻辑实例。

OSS 外部服务对象包括 DLF 元数据服务、权限和读写接口，DLF 将 OSS 目录数据识别为表的结构，MC 可以读取 DLF 探查的 OSS 元数据，按照 DataBase、Table 层次读写 OSS 数据。

Hadoop 外部服务对象包括 Hive 的 HMS 元数据存储，Kerberos 认证信息、HDFS 读写服务，可以读取 HMS 的 Hive 表元数据，将 Hive 的 DataBase、Table 映射为 MC 相同结构的外部项目。

以上的元数据自动映射方法免去了逐个表创建外表的繁琐工作，不冗余存储 DLF 或 HMS 的元数据，不冗余存储 OSS 或 HDFS 的数据，按照计算任务实时读取元数据，再访问数据。

3）日常联邦计算使用数仓中治理过的数据和 OSS、HDFS 中的外部数据直接关联分析，或从 OSS、HDFS 向数仓同步数据，实现湖

仓数据灵活流转。

4)基于数仓加工的数据，通过交互式分析引擎 Hologres 直读 MC 数仓 Pangu 文件，进行高效灵活的 OLAP 分析。

5）湖中的元数据信息通过外部服务自动映射，数仓元数据生产链路按周期抽取、同步给 Dataworks 数据地图，实现从湖倒仓的全链路血缘。

6）外部引擎如果需要数仓的数据，还可以通过 MC 的开放数据存储 MaxStorageAPI 和面向 Spark、Flink、Presto 等优化的 Connector 获取数据，保持了数据架构的灵活性和开放性。

4. 价值与效果

数仓向外部、低价值密度、弱管理的数据平台延伸管理能力是湖仓一体的一个基本形态，虽然弱于仓内强管控带来的完整数仓能力，但是也对湖数据管理能力进行了增强和统一。基于 MC+DLF+EMR+OSS 的湖仓一体架构具体效果有:

1）**实现湖仓统一管理。**通过自动映射元数据和数据读写能力简化了人工参与外部系统元数据在数仓中重新定义的工作和对联邦数据源的兼容适配工作。数据流转不需要创建外表或配置 ETL 流程，研发效率提升 50%以上。

2）**实现湖仓数据分层存储。**EMR 存储原始数据，数据中台对数据湖数据抽取构建 ODS 和 DW 层存储在 MC 上，其他引擎消费 ADS 层，数据存储比预期降低 50%。

3）**实现数仓高性能计算。**MC 数仓对外表的计算效率虽然低于

内表，但是湖到仓的数据抽取只需要执行一次，后续仓内计算效率是 Hive 的 5 倍以上，仓内建模加工效率也是 Spark 的 2 倍左右。

大数据技术标准推进委员会

地址： 北京市海淀区花园北路 52 号

邮编： 100191

邮箱： TC601@CCSA.org.cn

网址： www.tc601.com