

中国云原生数据湖应用洞察 白皮书

©2022.4 iResearch Inc.



概念界定：数据湖是面向大数据场景的创新解决方案，采用了与传统数仓不同的设计架构，具有「数据多源异构、统一存储管理、多范式计算、schema后置和应用广泛」的特性。云原生是数据湖未来部署的必然形态，具有「建立统一数据资产、低成本使用基础资源、高性能计算体验升级和敏捷创新赋能」的核心价值。



市场现状：数据变革、企业数字化转型、投融资、政策支持将持续加速释放云原生数据湖的应用需求。2020年云原生数据湖市场规模（含生态）达124亿，预计未来三年将以39.7%的复合增长率快速扩张。

竞争格局：中国云原生数据湖还处于发展的早期，能够提供整体解决方案的独立厂商还较少，市场较为集中，竞争主要围绕头部云厂商展开。以营收口径核算，2020年云厂商在中国云原生数据湖市场（不包含生态支持部分）的份额达到了82.4%。



应用现状：现阶段，云原生数据湖主要应用于泛互联网行业（40.7%）及传统行业的互联网场景（泛政务、金融、工业、医疗、汽车等），未来将向更多具有大数据和高价值属性的行业拓展。

选型建议：企业在布局数字化转型时，面对多元且快速迭代的业务需求，一方面需建设统一的数据底座，另一方面需关注DT能力的开放性、敏捷性和创新性。在选型云原生数据湖时，除内部能力评估外，还需要考虑服务商的服务半径和发展路径。



趋势展望：在云原生与大数据背景下，云原生数据湖成为企业智胜未来的新一代生产力工具，市场即将迎来爆发期。尽管数据湖与云和大数据天然契合（海量、弹性、简单、敏捷），但在具体业务场景落地中，仍有许多实际问题需要解决。未来，云原生数据湖厂商需与开发者、ISV和SI共同努力，在企业级生产环境中不断探索，生态共赢驱动云原生数据湖解决方案日臻完善。

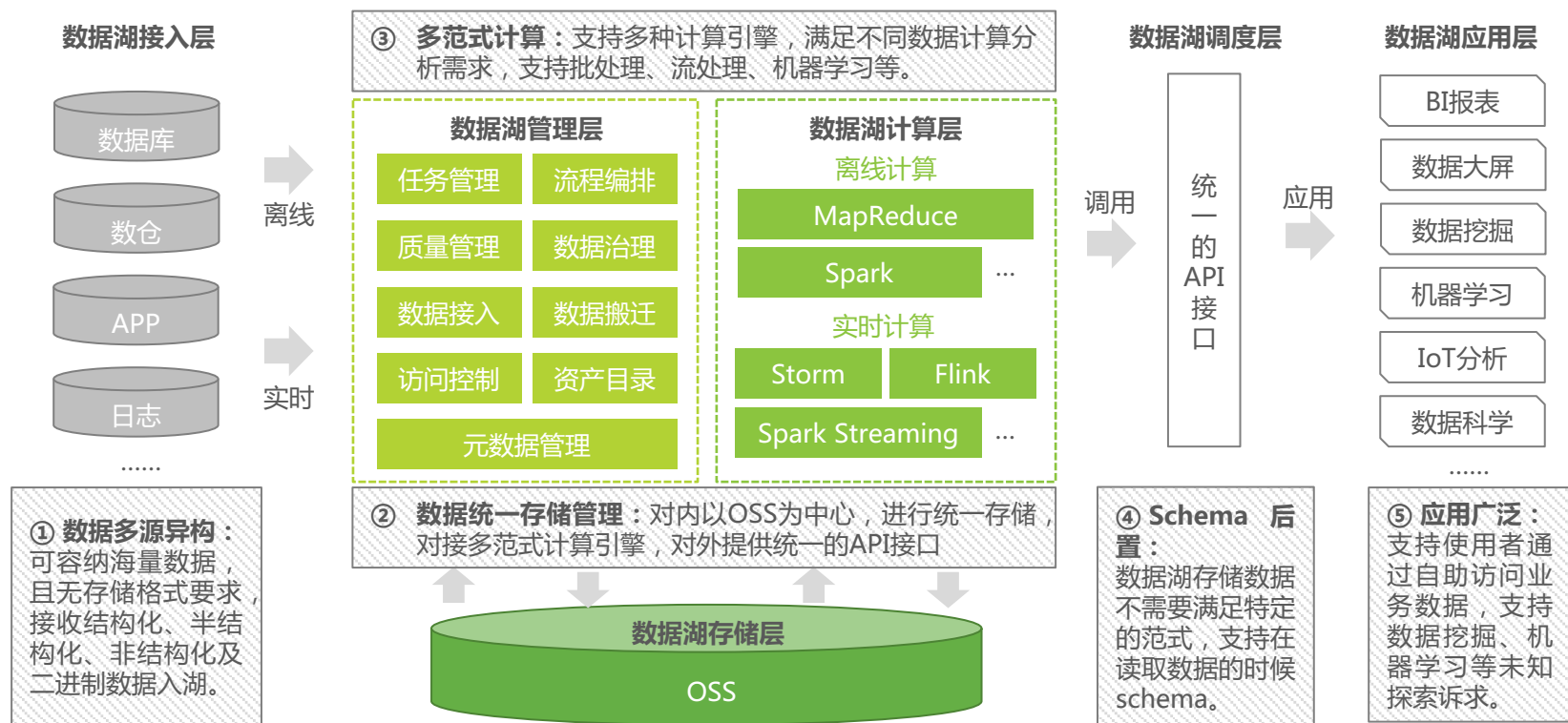
云原生数据湖概念界定	1
云原生数据湖市场现状	2
云原生数据湖竞争分析	3
云原生数据湖行业应用与最佳实践	4
云原生数据湖选型建议与典型企业	5
云原生数据湖发展趋势	6

数据湖的定义

数据湖是面向大数据场景的创新解决方案

早期，业界和用户多把数据湖定义为一个储存原始格式数据的系统，可容纳结构化、半结构化、非结构化及二进制的数据。随着大数据技术的融合发展，数据湖的边界不断扩展，内涵也发生了变化。数据湖开始汇集各方面技术，逐步演进成为集多源异构数据统一储存、多范式计算分析及统一管理调用的大数据综合解决方案。它可以更加高效率低成本地管理海量多源异构数据，打通数据孤岛，释放数据价值，助力新时代下各行业企业的数字化转型。

数据湖典型构架及特性



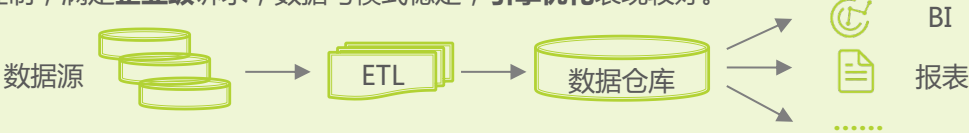
来源：艾瑞咨询研究院自主研究及绘制。

数据湖 vs 数据仓库

诞生背景、设计思路及使用场景各不相同

数据仓库是诞生于数据库时代，应企业分析诉求而生的数据产品，它的核心思路是把数据库中的数据进行一定格式转换后，定时地复制至另一个库里做列式存储，从而满足企业查询和数据分析的诉求。随着互联网的发展，数据量暴增，非结构化数据越来越多，企业业务变化越来越快，传统数据仓库无法适应大数据和现代化企业对于实时、交互式分析等方面的诉求。随之，数据湖诞生。它选择了“前松后紧”的设计思路，初始化阶段放弃严格的模式，后置schema，从而获取更强的灵活性；同时通过统一存储管理和计算优化来保证数据的一致性和性能。

数据湖与数据仓库对比

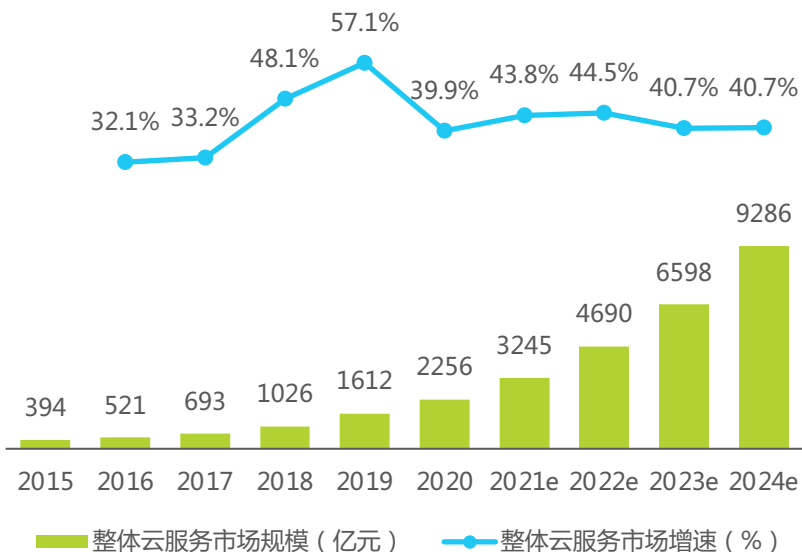
	 数据源	 数据处理	 适用场景	 性价比
 数据仓库	支持处理过后的结构化/半结构化数据；来自业务系统	写时建模 (Schema-on-write)	传统行业，以及企业的稳态业务； 数据量少，数据结构化，稳定可预测，对 执行实时性要求不高	建设成本低 扩容成本高
高度监管与严格事前控制，满足企业级诉求；数据与模式稳定，引擎优化表现较好。 				
 数据湖	支持未经处理的结构化/半结构化/非结构化数据；来自IoT设备、 Web、APP和业务系统等	读时建模 (Schema-on-read)	泛互联网行业以及传统行业的互联网场景； 海量数据，迭代速度快，需要实时分析	扩容成本低 建设成本高
可针对特定业务需求进行重新配置，灵活性和可扩展性较强。 				

来源：艾瑞咨询研究院自主研究及绘制。

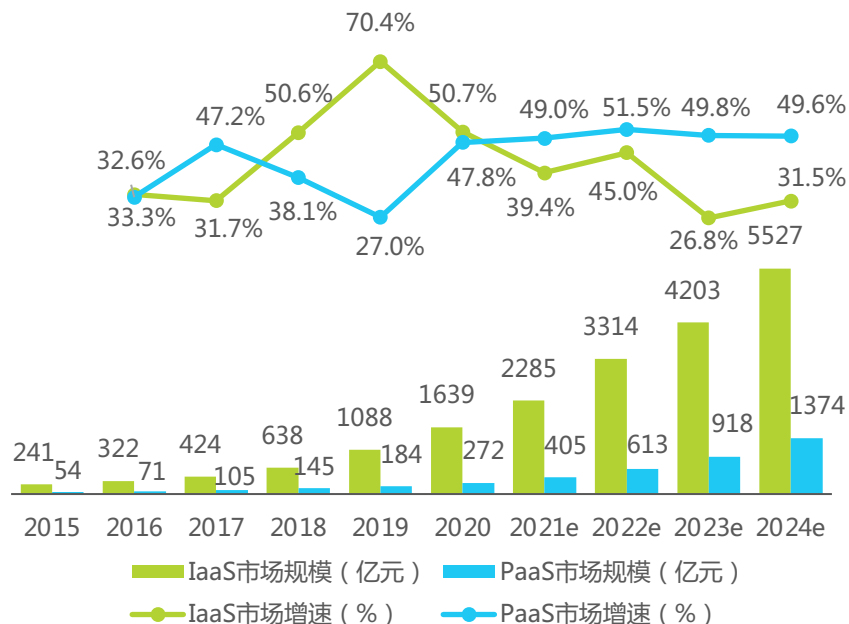
云原生部署是数据湖未来的必然形态

近年来，在数字经济的背景下，互联网行业及传统企业加速云化转型，中国整体云服务市场的规模逐年扩增，云成为新一代IT基础设施已经成为不争的事实。其中，企业云化转型的深入以及用云思维的转变，驱动了PaaS市场份额的增长，基于云的能力创新已成为基础云发展新的增长引擎。云特有的“池化、弹性、成本、敏捷”等优势让数据层与应用层的很多设想得以实现，拥抱云原生成为数据湖乃至大数据的必然选择。

2015-2024年中国整体云服务市场规模及增速



2015-2024年中国整体IaaS和PaaS市场规模及增速



来源：艾瑞《2021年中国基础云服务行业发展洞察》，艾瑞咨询研究院自主研究及绘制。

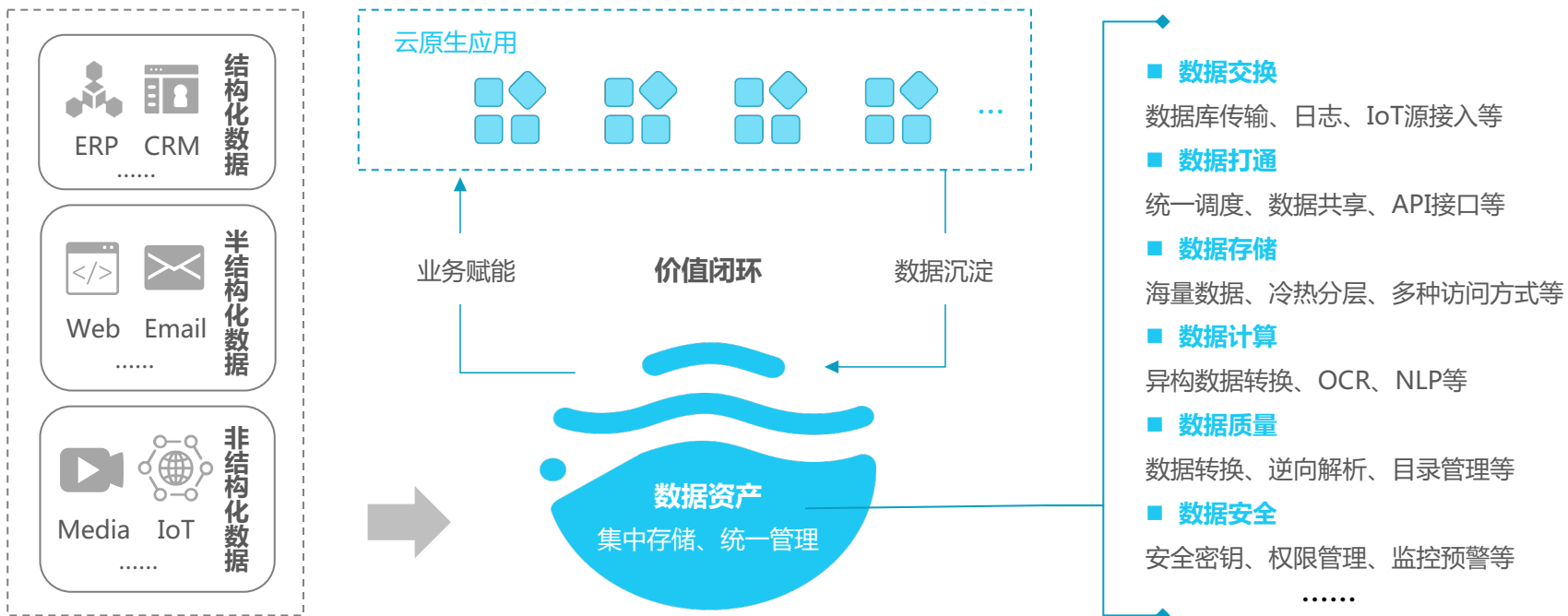
来源：艾瑞《2021年中国基础云服务行业发展洞察》，艾瑞咨询研究院自主研究及绘制。

云原生数据湖核心价值一：数据资产

集中存储、统一管理，建立高质量的数据资产

随着数字化转型进入深水区，“数据”已经成为企业的核心生产要素，打通各部门、各应用系统，建立企业级的统一数据资产已经成为业内的共识。基于云上的集中存储和数据湖，企业可以更丝滑地实现数据多源聚合，对内外部数据进行全生命周期的管理，从而沉淀为数据资产，赋能业务应用，释放数据价值。同时，基于云原生数据湖部署的云原生应用天然可以实现数据的无界流动，数用一体为企业打造了高效的价值闭环。

基于云原生数据湖的统一资产建设



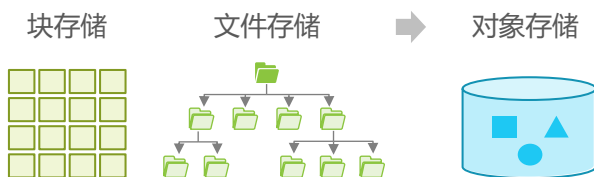
云原生数据湖核心价值二：低成本

通过云的方式，更低成本地使用存储和计算资源

云原生数据湖是基于云环境构建的低成本大数据解决方案。于存储上，云原生数据湖使用对象存储，实现了无限扩容（理论上）和更低的价格，同时云上统一存储也简化了之后数据调用的复杂度；于计算上，云原生数据湖采用计算存储分离的架构，让计算节点和存储节点可以分别弹性伸缩，避免了存算需求不同造成的浪费；于用云策略上，云原生数据湖通过 Serverless 的模式，根据请求量自动进行毫秒级的弹性扩容，解决波峰资源短缺、波谷资源浪费的问题，实现最小单元的成本最优。

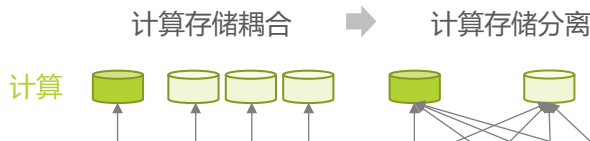
云原生数据湖成本优化剖析

存储成本 OSS



不同于直接操作物理磁盘的块存储，或基于文件路径访问的文件存储，对象存储通过唯一标识符（Key）映射寻址，存取都非常灵活和简单。这种方法对在云计算环境中自动化和简化数据存储都大有裨益，体现在用户侧即表现为理论上无限的扩容可能性和更低廉的存储成本。

计算成本 计算存储分离



随着移动互联网、产业互联网、5G的发展，个人端和企业端产生的数据量爆发增长。在早期大数据分析的架构下，计算资源和存储资源是紧耦合的，只能同步扩容，这造成了计算资源的过剩。存算分离后，计算节点和存储节点都

预览已结束，完整报告链接和二维码如下：

https://www.yunbaogao.cn/report/index/report?reportId=1_41142

