

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。

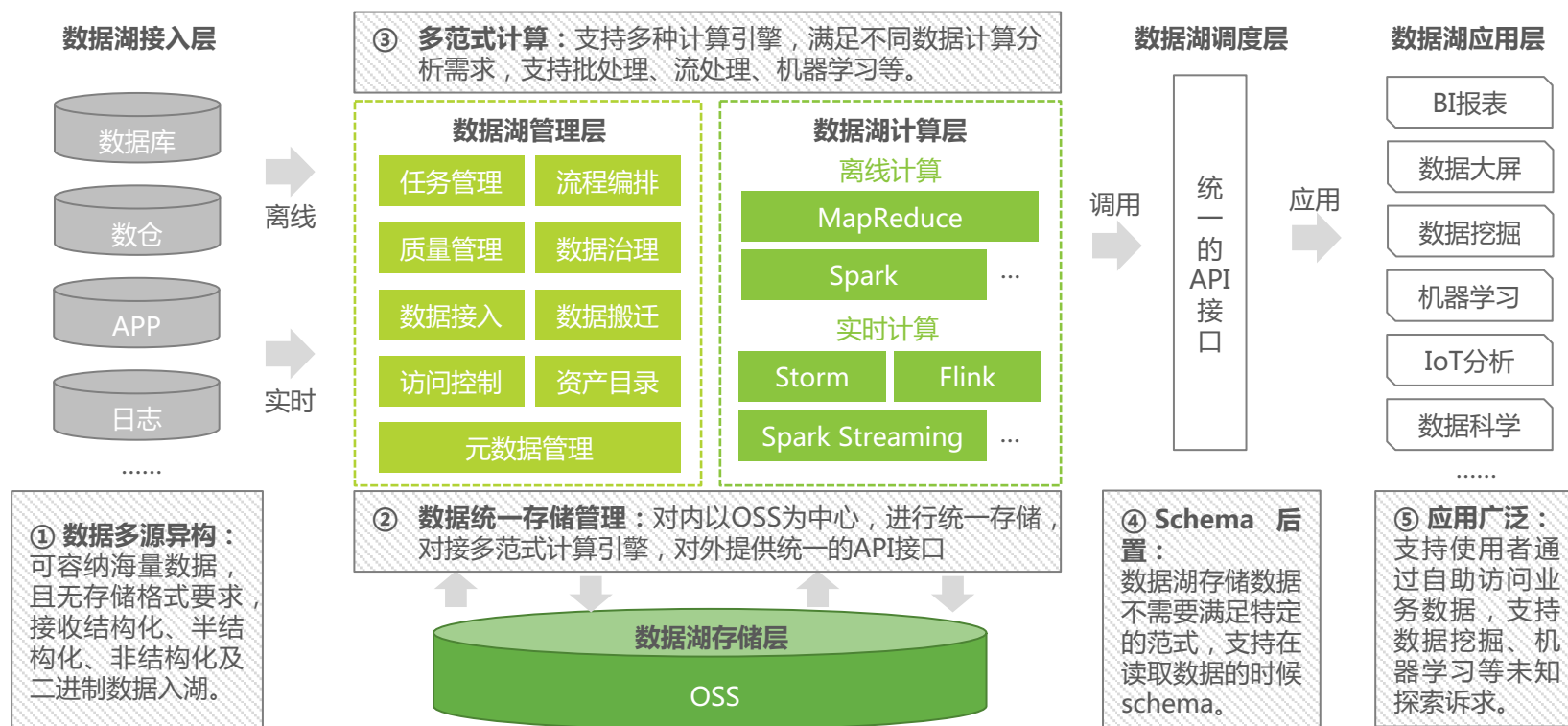
云原生数据湖概念界定	1
云原生数据湖市场现状	2
云原生数据湖竞争分析	3
云原生数据湖行业应用与最佳实践	4
云原生数据湖选型建议与典型企业	5
云原生数据湖发展趋势	6

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库。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

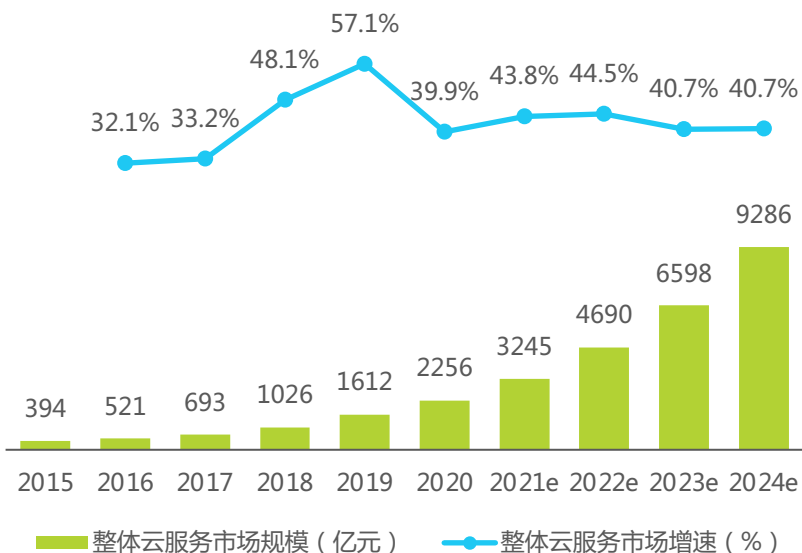
	 数据源	 数据处理	 适用场景	 性价比
 数据仓库	支持处理过后的结构化/半结构化数据；来自业务系统	写时建模 (Schema-on-write)	传统行业，以及企业的稳态业务； 数据量少，数据结构化，稳定可预测，对 执行实时性要求不高	建设成本低 扩容成本高
高度监管与严格事前控制，满足 企业级 诉求；数据与模式稳定， 引擎优化 表现较好。				
				
 数据湖	支持未经处理的结构化/半结构化/非结构化数据；来自IoT设备、 Web、APP和业务系统等	读时建模 (Schema-on-read)	泛互联网行业以及传统行业的互联网场景； 海量数据，迭代速度快，需要实时分析	扩容成本低 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

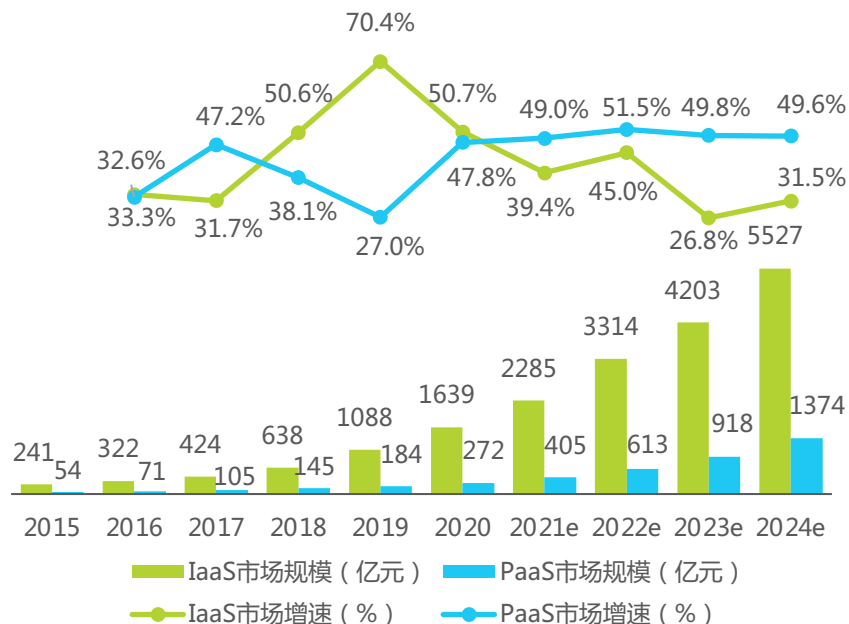
云原生部署是数据湖未来的必然形态

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。

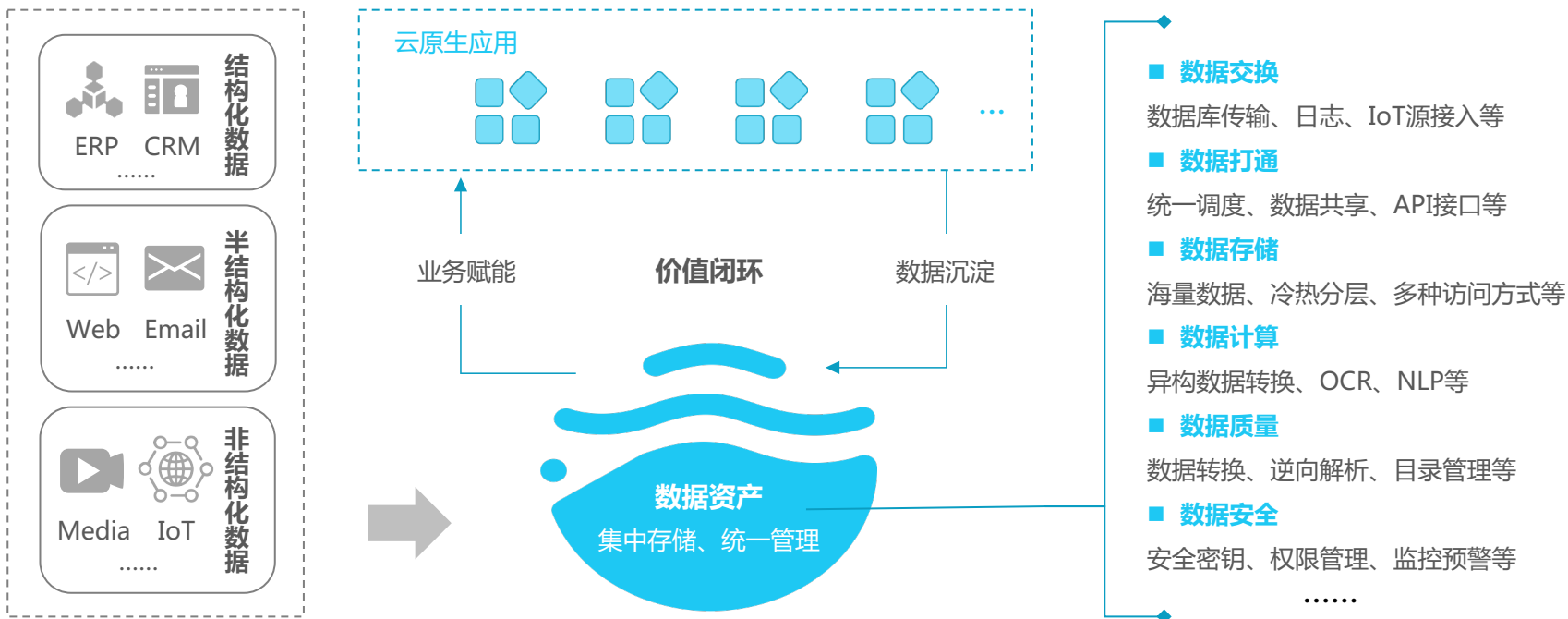
来源：艾瑞《2021年中国基础云服务行业发展洞察》，艾瑞咨询研究院自主研究及绘制。

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

随着数字转型化进入深水区，“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



来源：艾瑞咨询研究院自主研究及绘制。

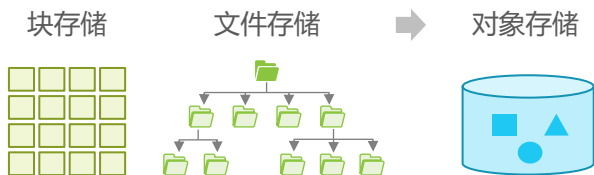
云原生数据湖核心价值二：低成本

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。

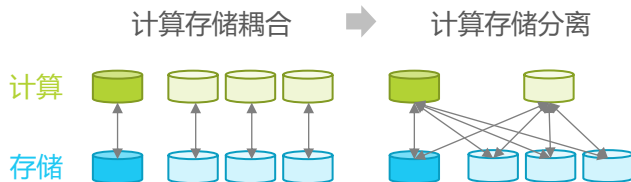
云原生数据湖成本优化剖析

存储成本 OSS



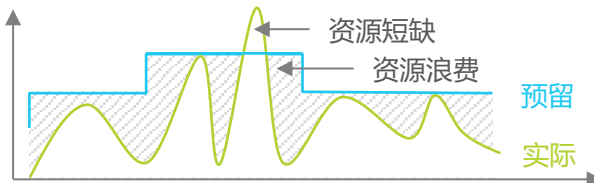
不同于直接操作物理磁盘的块存储，或基于文件路径访问的文件存储，对象存储通过唯一标识符（Key）映射寻址，存取都非常灵活和简单。这种方法对在云计算环境中自动化和简化数据存储都大有裨益，体现在用户侧即表现为理论上无限的扩容可能性和更低廉的存储成本。

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可以按需弹性扩容，大大降低了计算的成本。

用云成本 Serverless



现阶段定时等云资源调用机制具有一定的滞后性，为了保证高可用，企业往往选择采取冗余的伸缩策略，这造成用云成本的上升。Serverless模式下，资源消耗随着应用程序的需求（请求数量）变化自动扩展或缩减，计费精确到毫秒级，大大降低了企业数据湖用云成本高企的问题。

云原生数据湖核心价值三：高性能

云湖共生，带来大数据应用的高性能体验

数据湖“统一→简单、松耦合→弹性、敏捷→探索”的设计思路与云计算天然契合，当数据湖以云原生的方式部署时，其强大的性能优势可以被最大化释放。一方面，数据湖上云后可以享受云本身带来的性能提升，如高可用、弹性、敏捷等；另一方面，数据湖在云原生的环境中可以做更多性能优化的工作，如丰富的上下文带来的分析加速，流批融合带来的实时数据价值释放，一站式数据管理方案带来的安全和质量改善等。

云原生数据湖性能优化剖析

01 On Cloud 本身带来的性能提升

高可用

相较自建IDC，云环境拥有更多的资源冗余，一节点发生故障能无缝切换到其他节点，从而对企业侧体现为高可用，确保了业务的连续性。

弹性

云计算具备动态扩充性与可负担性，可以解决海量业务带来的吞吐和IO性能瓶颈，满足大数据分析所需资源的庞大规模与突发性质的需求。

敏捷

云让企业得以从重复、复杂的底层IT工作中释放出来，同时其模块化、松耦合的敏捷架构有利于数据产品的快速迭代、部署、运维和创新。

02 In Cloud 更进一步地性能优化

加速

一方面，云原生数据湖提供了比以往更丰富的上下文，有助于加速分析实验；另一方面，它统一了流式处理和批式处理，可以为企业提供更实时的数据价值体验。

多范式

云原生数据湖基于云环境统一了企业数据资产和多范式计算引擎，从而可以支持企业对任何数据类型执行任何分析。同时其可扩展的架构也为企业使用AI进行探索做好了准备。

安全

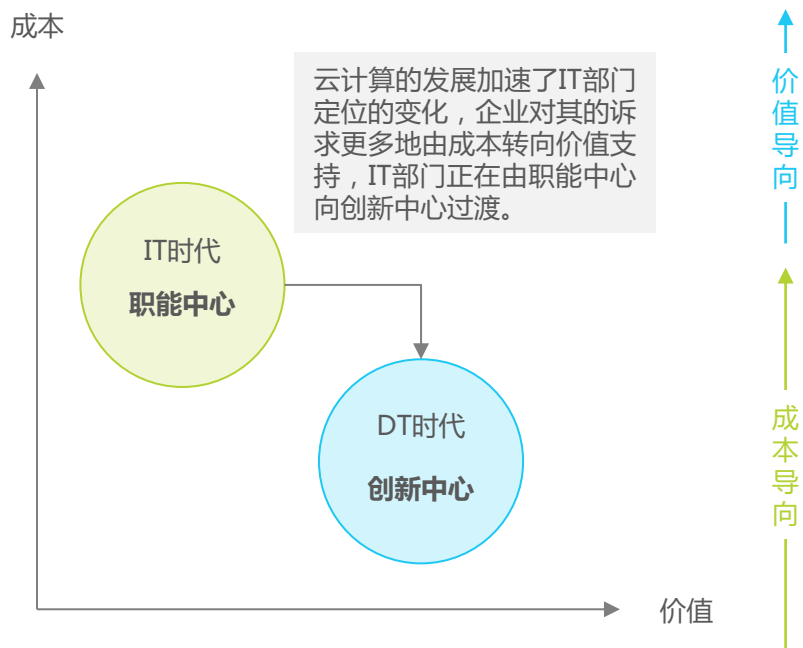
云原生数据湖提供了简单、强大的数据管理解决方案，以全保真的方式存储任何类型或数量的数据，有助于企业加强安全和治理。

云原生数据湖核心价值四：敏捷创新

重塑IT部门定位与价值，赋能业务应用敏捷创新

云服务重塑了IT产业的分工和企业IT部门工作的内容和方式，企业IT部门越来越少地关注复杂的底层技术，转而向应用创新聚焦，充分释放其业务赋能价值。通过统一对象存储、多引擎兼容、数据智能管理，云原生数据湖基于云的环境进一步释放了企业IT的生产力。IT部门无需再关注基础资源和数据层的大多问题，如存储扩容、计算优化等，可以将更多的精力放在业务支持、应用创新上，实现真正的数据驱动企业发展。

企业IT部门定位变化



云原生数据湖的应用创新价值

应用层

将云原生数据湖作为企业大数据的解决方案，可以更进一步地屏蔽底层的复杂性，聚焦于业务服务。基于弹性的IT基础资源和柔性的数据资产，IT可以更敏捷地进行应用创新。

计算层

在云原生的环境下，企业可以在统一的控制台上简单地（写SQL一种语言即可）进行多范式计算，根据业务需求和数据属性自动/半自动地选择适合的计算引擎，无需IT部门再花费额外的学习成本去进行计算优化。

存储层

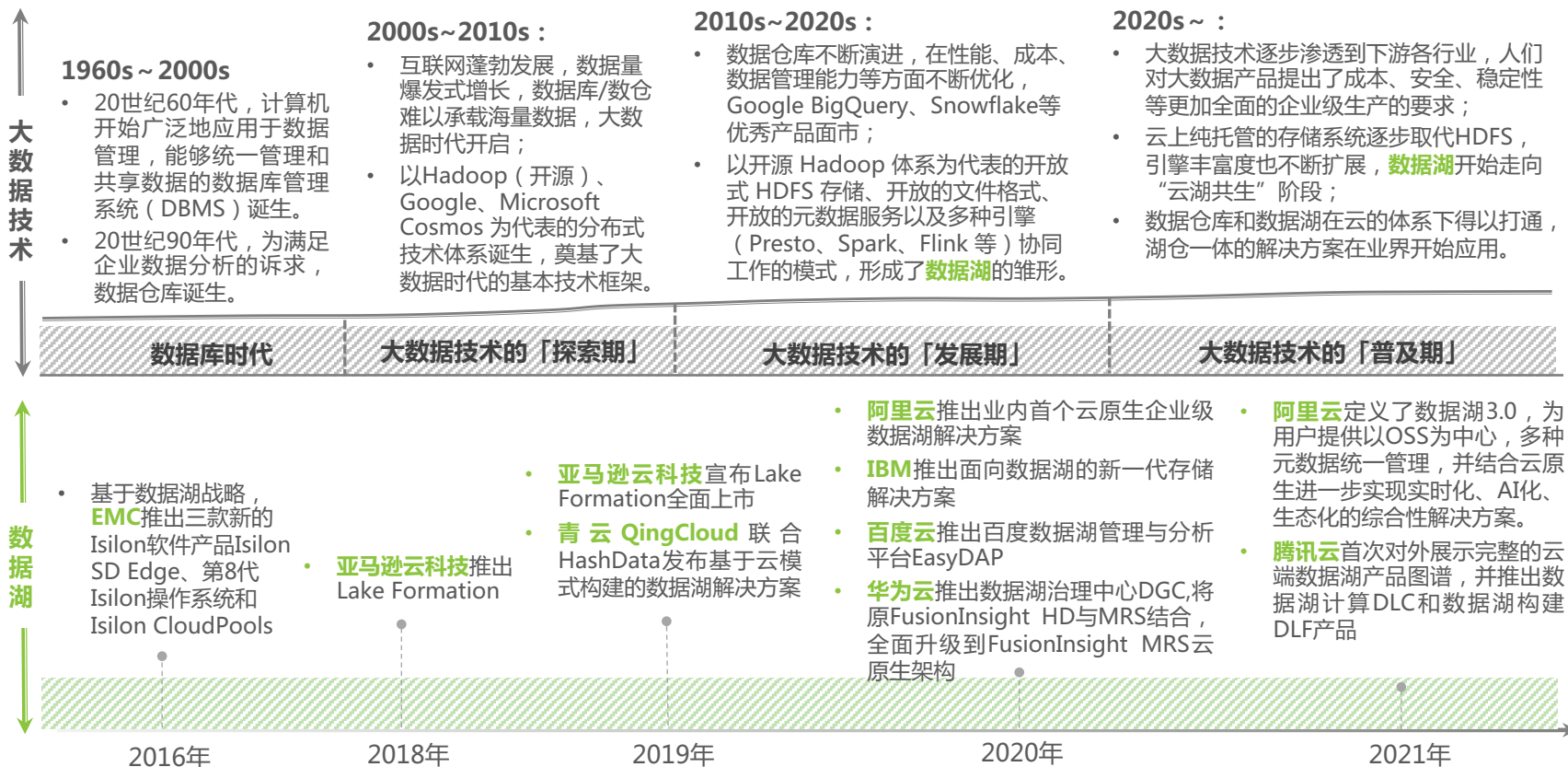
基于云原生对象存储的方式，企业无需担心数据增长带来的扩容问题，无需关注数据存储的物理位置，只需要将云当作是一个无限扩展、简单存取、弹性伸缩的“网盘”即可。

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

产品随市场需求不断演进，国内数据湖尚处于发展初期

中国数据湖技术正在逐年发展及突破，公有云厂商及其他行业厂商纷纷在做尝试。但目前数据感知收集及归类清洗方面存在壁垒和难度，数据湖建模经验不足，因此我国数据湖市场整体发展处于初期阶段，未来发展空间广阔。

中国云原生数据湖行业发展历程

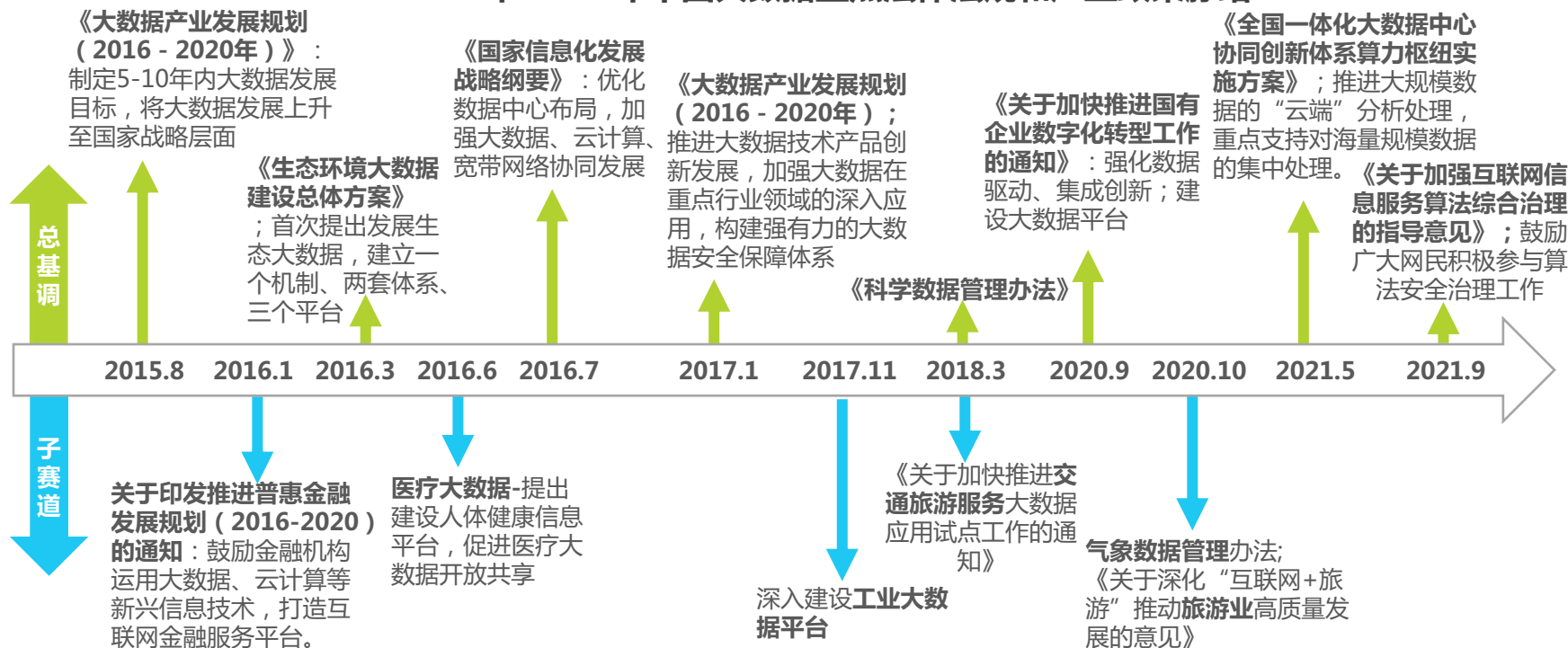


来源：公开资料，专家访谈，艾瑞咨询研究院自主研究及绘制。

法律法规不断落地，推动大数据产业走向成熟

2015年出台的《促进大数据发展行动纲要》呈现“一体两翼一尾”的格局，首次将大数据发展提升至国家战略层面，奠定了大数据未来发展的总体基调。2021年5月印发的《全国一体化大数据中心协同创新体系算力枢纽实施方案》提出加快建设全国一体化大数据中心算力枢纽体系，同时加强对基础网络、数据中心、云平台、数据和应用的一体化安全保障，提高大数据安全可靠水平。近五年间，国家出台多条产业政策及法规，不仅从方针上引领大数据产业高效、合规发展，同时也将该产业布局至政务、金融、工业、医疗、旅游服务、气象管理等多个细分领域。

2015年-2021年中国大数据重点法律法规和产业政策脉络

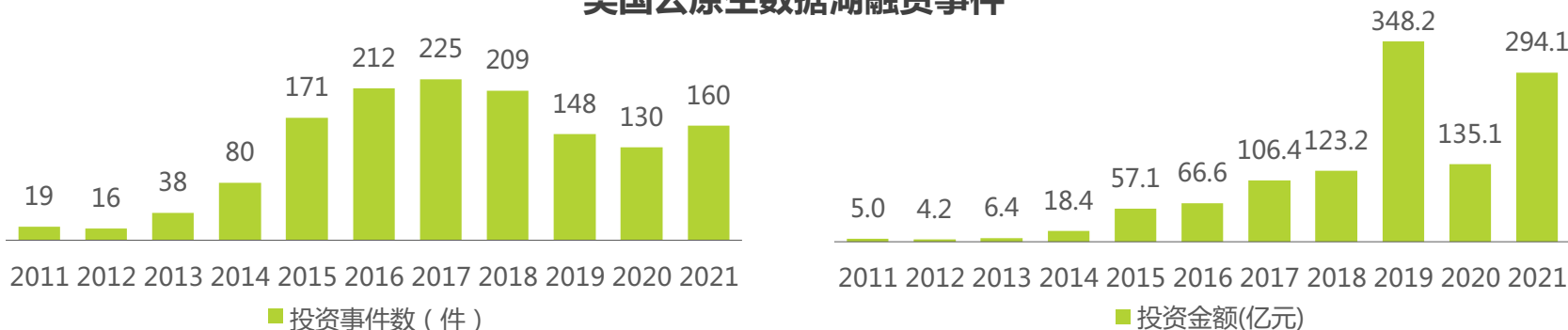


来源：中国政务网国务院政策文件库，艾瑞咨询研究院自主研究及绘制。

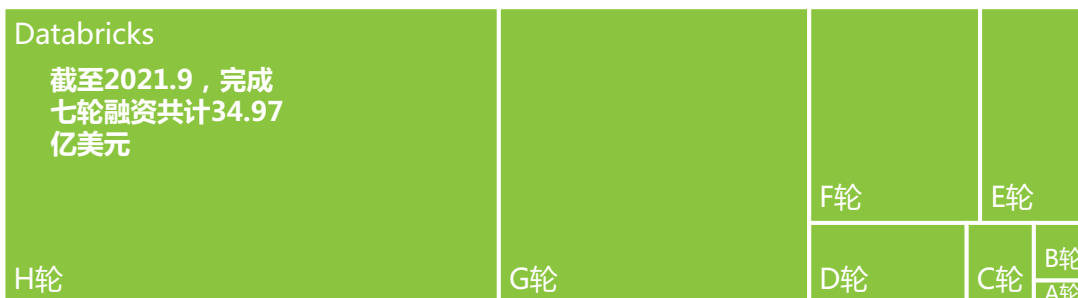
资本市场活跃，数据湖商业价值逐步凸显

据统计，近年来数据服务行业投融资事件数和金额整体呈上升趋势，并在2019年达到了巅峰，投资金额超过了300亿元人民币。2020年，受到疫情等外部因素的影响，投融资事件数和金额数均有所下降。但随着国内疫情的稳定和经济的回暖，2021年，数据服务行业的投融资再次展现出上升的态势。放眼全球云原生数据湖市场，初创数据湖厂商Databricks、Upsolver等都获得了上亿美元的融资。该领域的资本市场活跃，数据湖的商业价值逐渐凸显。

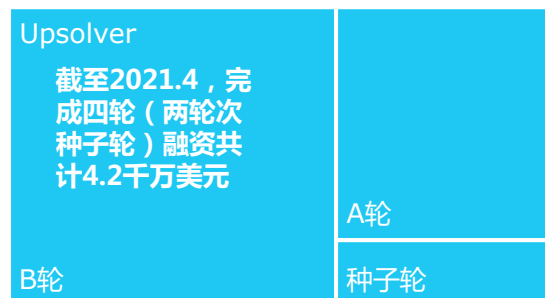
2011-2021年中国数据服务行业投资情况及美国云原生数据湖融资事件



Databricks



Upsolver



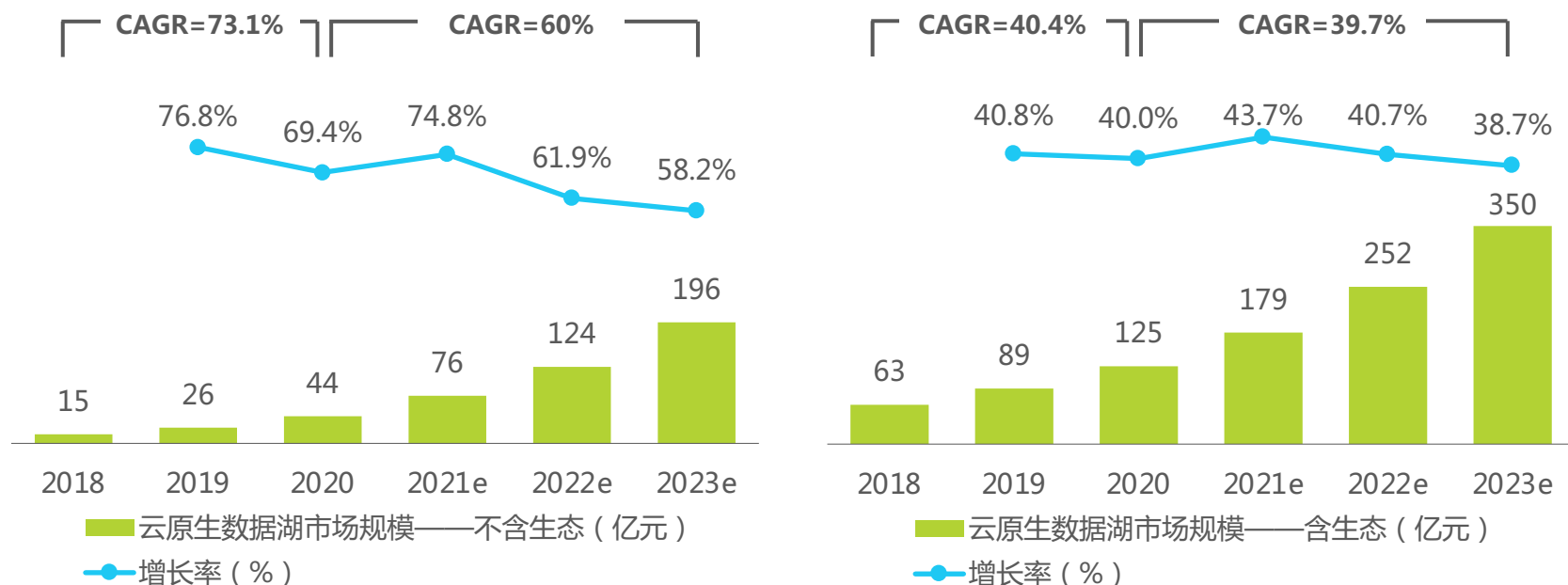
来源：IT桔子，Crunchbase，艾瑞咨询研究院自主研究及绘制。

中国云原生数据湖市场规模

2020年规模达124亿，预计未来三年维持39.7%的快速增长

据艾瑞统计，中国云原生数据湖2020年整体规模达124.8亿元。目前行业正处于初期发展阶段，由于国家政策利好、互联网技术高速发展的驱动、企业数字化转型加速等因素，预计中国云原生数据湖市场未来三年会以39.7%的复合增长率快速发展。

2018-2023年中国云原生数据湖市场规模及增速



注释：云原生数据湖市场规模——不含生态统计口径为2020自然年全年各厂商在中国内地（不含港澳台）销售云原生数据湖解决方案的营业收入，合同签署地点和交付地点都位于中国内地区域。不包含云原生数据湖组件（包括存储、计算、管理及调度层）发生于其他解决方案（例：数据仓库）的营收，不包含云原生数据湖生态支持厂商部分。

注释：云原生数据湖市场规模——含生态统计口径为注释1中包含云原生数据湖组件（包括存储、计算、管理及调度层）发生于其他解决方案（例：数据仓库）的营收，且包含云原生数据湖生态支持厂商部分。

来源：艾瑞长期基础云服务数据监测，结合公开资料、专家访谈，根据数据测算模型，自主研究及绘制。

数据变革与企业数字化转型加速云原生数据湖的应用

IoT、移动互联网和5G的发展，带动数据量爆发，如何从数据海啸中挖掘数据价值成了企业亟待解决的难题。在此背景下，企业亟需新的大数据架构来处理数据，这为数据湖市场发展带来契机。互联网的发展加速了时代数字化发展，同时也深刻地改变了企业的业务模式。以“敏捷、创新、数据驱动”为导向的数字化转型需要新的生产力工具来打破数据孤岛、沉淀数据资产、完成数据价值反哺企业。云原生数据湖的各部分组件为数字化转型的每一阶段提供技术支持，完成“数”与“智”的融合。

云原生数据湖为企业数字化转型各环节提供技术支持



来源：艾瑞咨询研究院自主研究及绘制。

在数据治理、全链路、安全等方面仍待持续改进

从应用现状来看，数据湖在国内的落地还存在许多痛点。产品层面，数据湖的数据治理能力和全链路能力仍需进一步的加强，客户方更亟需智能化、一站式的解决方案；应用层面，云原生数据湖的行业认知和人才培养较为单薄，仍待市场的进一步培育。另外，近期安全隐私法律法规不断落地，企业主对云原生数据湖的安全监管也提出了更高的要求。

云原生数据湖应用的集中痛点

人才缺失

目前大数据、AI技术栈日新月异，企业缺乏专业人才。从企业内部来看，管理者对数据治理一知半解，若在没有深入梳理企业业务现状及需求的情况下盲目搭建数据湖、追求“大而全”的概念，可能导致数据湖落地效果不佳。

行业认知

尽管数据的价值属性已经获得业界的广泛共识，但是选择观望的企业依旧占据大多数，数据湖在认知和推广上仍然面临着多方面的挑战。

安全监管

随着企业数字化进入深水区，“数据”已经成为市场和企业的核心生产要素。数据湖的最大风险之一就是安全性和访问控制。大量数据可以在没有任何监督的情况下流入湖泊，一旦某些数据包含其他数据所没有的隐私和法规要求，将会有一定几率发生数据泄露或者遗失，后果不可估量。

全链路能力

现阶段国内可以提供全链路云原生数据湖服务的供应商较少，大多厂商仅提供数据湖组件的支持，因此下游需求企业只能采购多家供应商来满足自身从数据采集治理到分析可视化的需求。尤其是技术水平较弱的企业更为希望厂商可以提供全面的服务。

数据治理

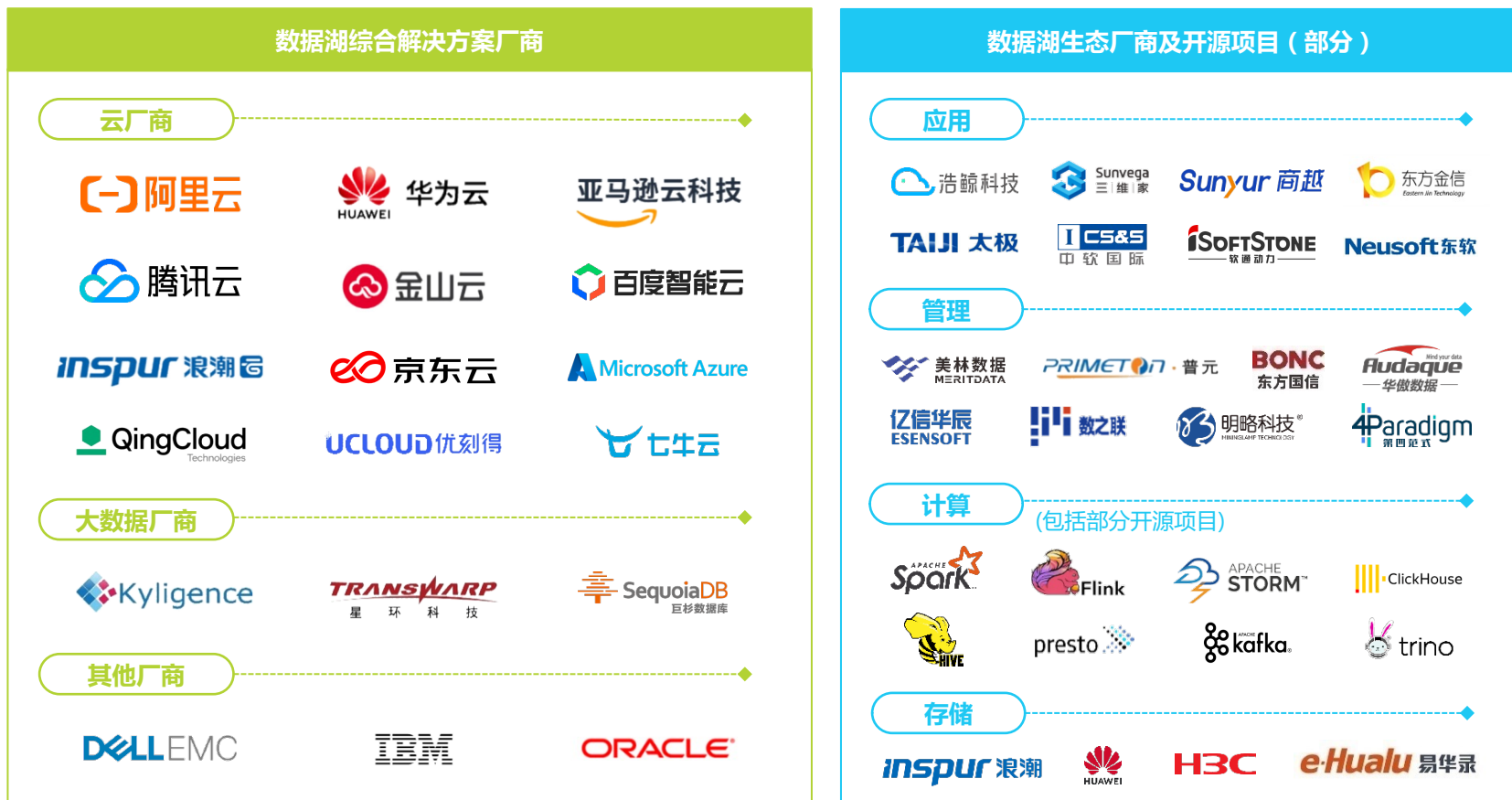
数据治理要求在目录中包含数据的分类、规则，若企业对于数据湖的掌控能力不足，会导致数据湖目录及整体构架设计不良、湖内数据未得到充分归档或维护，容易形成数据沼泽。因缺少上下文元数据关联，数据沼泽就无法进行数据检索，致使用户无法有效分析和利用数据。



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

公有云厂商+生态厂商的市场格局初现

中国云原生数据湖产业图谱

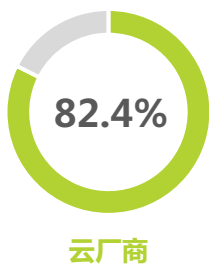


注释：此页主要表现云原生数据湖及其生态的布局情况，仅展示部分典型企业，图谱中所展示的公司logo顺序及大小并无实际意义。
来源：艾瑞咨询研究院自主研究及绘制。

先发优势，云厂商市占率达82.4%

整体来看，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。以营收为核算口径，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%。一方面，于先发优势上，云计算具有弹性算力支持、数据聚合的特性，与数据湖思路天然契合；另一方面，于布局实践上，“春江水暖鸭先知”，出于服务自身或互联网客户的动因，云厂商率先基于云原生进行了能力的整合，在对象存储、多范式计算、大数据管理等云原生数据湖核心技术上都更为成熟。

2020年中国云原生数据湖市场（不含生态部分）竞争格局



1. 基础资源支持

云基础资源池化、存算分离的特性，可以最大程度上弹性、低成本地支持数据湖的各种工作。

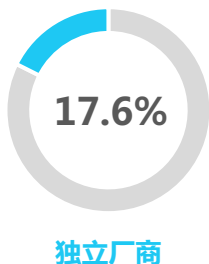
2. 数据聚合优势

基于云的形式，企业可以更丝滑地实现各系统相通，解决数据孤岛问题。

3. 能力统一调度

在云原生的环境下，企业可以以统一视角，更优雅地调用多种计算引擎。

国内市场环境复杂多变，在行业应用、客户服务等领域，云厂商还需要更多的生态厂商的补足。



1. 多云、混合云管理能力

独立厂商具有第三方中立性，可以支持多云部署管理，解决企业供应商绑定的后顾之忧。

2. 轻量与专注

与云厂商从云出发到数据服务的视角不同，独立厂商大多从数据服务出发，业务更加轻量与专注。

由于数据湖较其他大数据产品更强调“海量异构数据统一存储、多源数据统一管理、多计算引擎统一调用”的能力，故而对于第三方独立厂商而言，解决“海量存储、计算优化、生态建设”的成本都会更高，市场进入的难度也更大。

注释：独立厂商包括大数据厂商、软件厂商、以及其他提供云数据湖服务的IT厂商。

注释：此市占率统计口径为2020自然年全年各厂商在中国内地（不含港澳台）销售云原生数据湖解决方案的营业收入，合同签署地点和交付地点都位于中国内地区域。

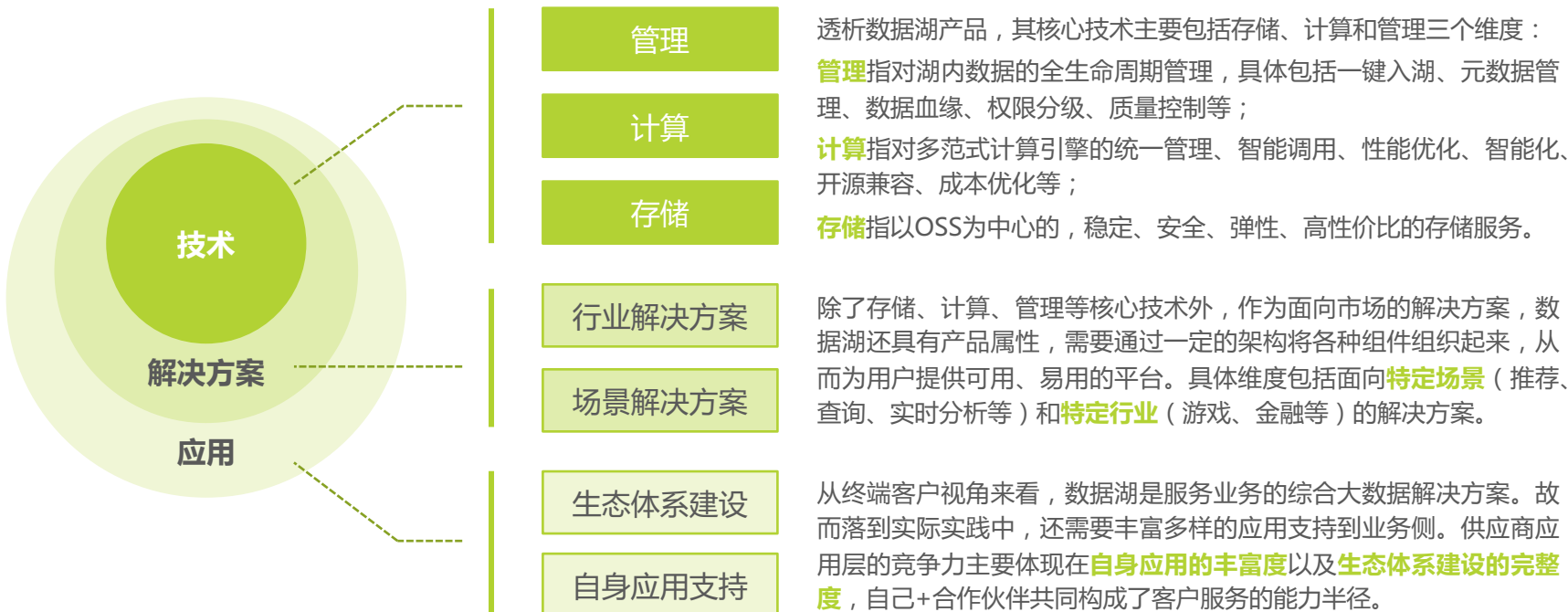
注释：此市占率统计口径不包含云原生数据湖组件发生于其他解决方案（例：数据仓库）的营收，不包含云原生数据湖生态支持厂商部分。

来源：艾瑞长期基础云服务数据监测，结合公开资料、专家访谈，根据数据测算模型，自主研究及绘制。

技术+解决方案+应用构成云原生数据湖的核心竞争力

云原生数据湖是一套完整的云上大数据解决方案，可以服务企业的多种数据诉求，其竞争要素可以归类为技术、解决方案、应用三层。在技术层，云原生数据湖需要具备稳定、高性价比的存储服务、多引擎兼容的计算优化服务以及全生命周期的智能化数据管理服务；在解决方案层，厂商需要贴近业务，面向特定场景和特定行业提供丰富、可落地的架构方案；在应用层，厂商需要通过生态或自建等方式提供更多的应用服务，不断扩大自己的服务半径，向终端客户展示更全面的能力。

云原生数据湖竞争分析框架



竞争要素一：技术

统一可靠存储+多元弹性计算+一站式智能管理

剖析云原生数据湖的核心技术，主要从存储、计算和管理三个维度去评估市面上的厂商。除了大数据产品通用的性能、可用性、安全及成本外，云原生数据湖还需要关注一些特定的竞争要素，如：存储层需要做前置的扩展性、性能和成本优化，以屏蔽硬件复杂性，支持多范式计算和大数据环境；计算层需要做多计算引擎优化和智能驾驶舱来简化企业使用流程；管理层需要支持多源湖外数据兼容和湖内数据全生命周期的一站式管理。

云原生数据湖核心技术

管理

兼容 | 一站式 | 安全

兼容：提供足够强大、丰富、高效（no-code）的连接器和转换工具，支持更多数据源的接入，支持更多种类的数据转换，满足各种场景诉求。

一站式：具备入湖、元数据、权限、血缘、质量、探索等数据湖所需的完备功能，提供一站式服务。

安全：支持数据任务看板、作业进度统计、日志审计、资源消耗统计、数据全链路展示、数字字典回溯追踪等功能，确保数据全生命周期的安全。

计算

多元融合 | 简单智能 | 成本优化

多元融合：可以兼容多种开源/商业计算引擎，满足企业数据处理的多种诉求，且进行了优化工作，使得多范式计算对客户侧表现为统一和简单。

简单智能：通过AI和Serverless，实现自动预配和管理计算资源，智能弹性伸缩工作负载以最大化资源利用率，简化运营运维工作，让团队可以专注于编程，不必管理服务器集群。

成本优化：资源自动伸缩叠加费用优化的批流引擎调用处理方案，让数据湖可以更为灵活地处理请求，在保证结果满意的前提下实现成本最优。

存储

稳定 | 扩展 | 简单 | 高效 | 性价比

稳定：具备成熟的物理冗余、传输校验、角色权限、安全加密方案，确保存储的最终稳定性。

扩展：数据湖承载的数据量每天都在持续增长，需要可以按容量灵活扩展的存储系统进行支持。

简单：面对应用对持久性、可用性和延迟的多样化要求，以及物理硬件复杂性，需要从存储层就着手进行优化，减少处理硬件资源复杂性的相关难题，使各应用程序都可以轻松获取和使用所需存储。

高效：面对海量数据，需要智能的冷热分层策略实现资源的均衡配置，提高服务效率、降低延迟。

性价比：海量数据带来对存储资源的大量需求，需要配置以相应的成本优化方案。

来源：艾瑞咨询研究院自主研究及绘制。

竞争要素二：解决方案

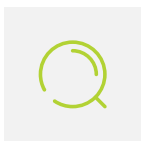
多场景挖掘+行业贴身服务

从市场现状来看，云原生数据湖并不是一个标准化的产品，而是一套松耦合、多模块、服务化的解决方案，在具体应用中还需要根据企业具体需求，进行组件调整和架构设计。因而，客户在选型采购时，除了关注厂商的技术实力，还会关注其解决方案的成熟度。具体评估维度包括2个方面：①厂商对数据湖典型适用场景的提取能力和方法论总结，这可以帮助项目更快速地实施；②厂商对具体行业业务的理解以及相应的实施思路，这可以帮助企业切实解决其痛点。

云原生数据湖典型解决方案

场景维度

海量数据交互式查询



在一些业务环节，如广告投放、用户运营、周报/月报等，需要对来自各个渠道的实时数据和历史数据进行交互式查询分析。云原生数据湖架构下，企业能够调用分布式的查询引擎，更加灵活、快速、准确的进行查询分析，支持业务决策。

企业级大数据治理



随着互联网的发展，企业内部积累了大量的数据，数据存储的成本愈发高昂，但数据价值却难以被全部释放。数据湖冷热分层的存储方案可以帮助企业将数据低成本的“存下来”，统一管理的架构让数据可以随时“用的到”，多种计算引擎兼容让数据可以“用的好”。

机器学习与AI探索



在风控、推荐、预测等场景，往往会需要机器学习加以支持。然而机器学习与AI探索会消耗GPU等大量的算力资源。云原生数据湖Serverless按需付费、自动扩容的方案降低了企业进行机器学习的TCO；同时schema后置的架构也让未知探索变得更加灵活。

行业维度

社交



基于移动互联网的社交平台近年来快速发展，短视频、直播、图片、资讯等构成了其服务的内容，大量非结构化数据的审核处理、实时分析、精准推荐为其带来挑战与机遇。云原生数据湖冷热分层存储、上下文关联分析、实时推荐的功能大幅提升了其内容创新和用户运营的效果，并降低了成本。

游戏



5G、云、社会娱乐方式、出海等因素共同驱动了游戏产业的快速发展。游戏大数据需要更智能、灵活、低成本的数据湖解决方案来进行实时动态监测、用户画像和运营分析，从而降低获客成本、改善游戏体验、留存现有玩家、提升付费转化率。

汽车



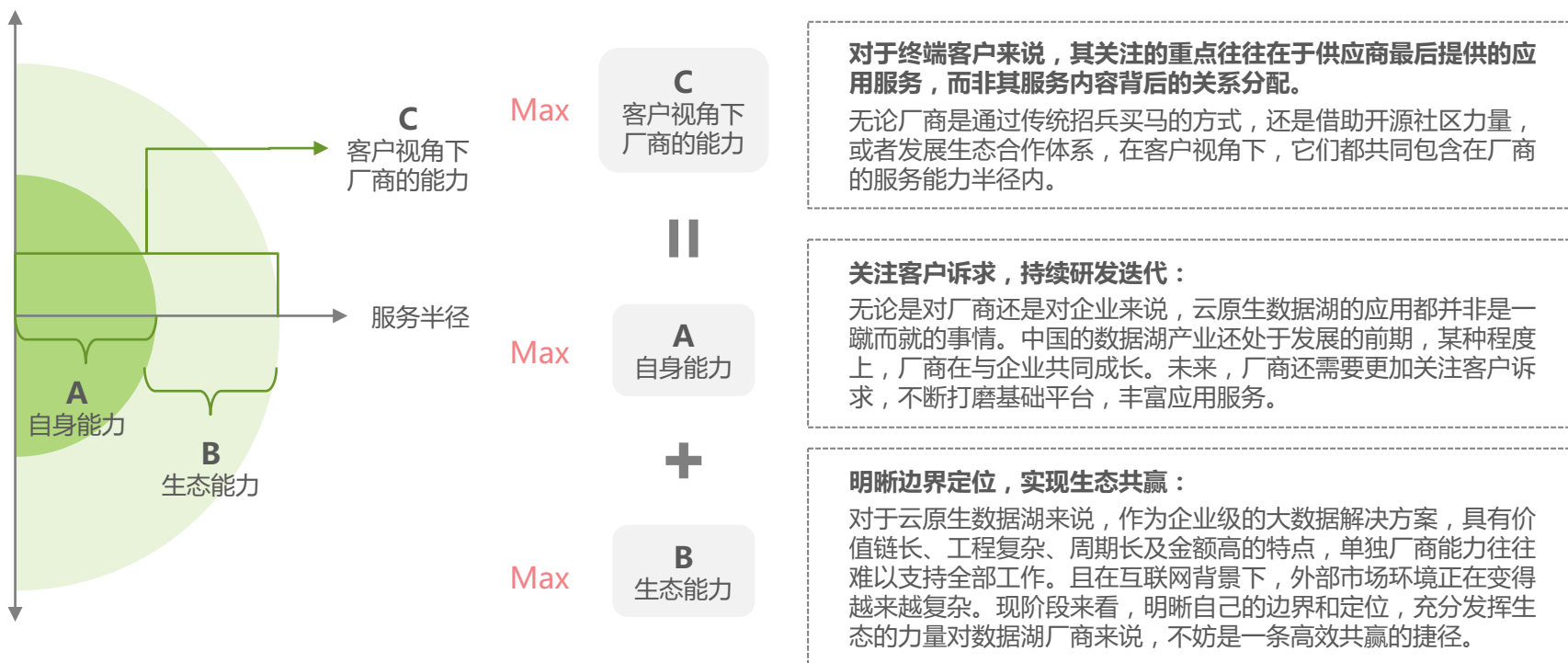
汽车正在成为未来生活的第三空间，车联网产业进入快车道，新型应用蓬勃发展，产业规模不断扩大。云原生数据湖可以实时地接收和存储车联网PB~EB级的数据，以低成本的方式进行资源调度，支持自动驾驶、智能交通等业务需求。

竞争要素三：应用

自研深耕+生态补充

就当今市场环境来看，大多厂商和企业都把数据湖定位为数据基座，但就实际使用来说，基座还需要配合具体的应用，才能真正地赋能业务。出于发展初期或产业分工的原因，现在云原生数据湖综合解决方案厂商还不具备提供完备应用服务的能力。且由于市场环境复杂，企业需求多变，在未来一段时间内，也很难有“一应俱全”的厂商出现。故而，除了关注客户需求，不断自研深耕外，厂商还需要通过生态建设，补足自己的服务半径，增强企业服务的竞争力。

云原生数据湖厂商的服务半径



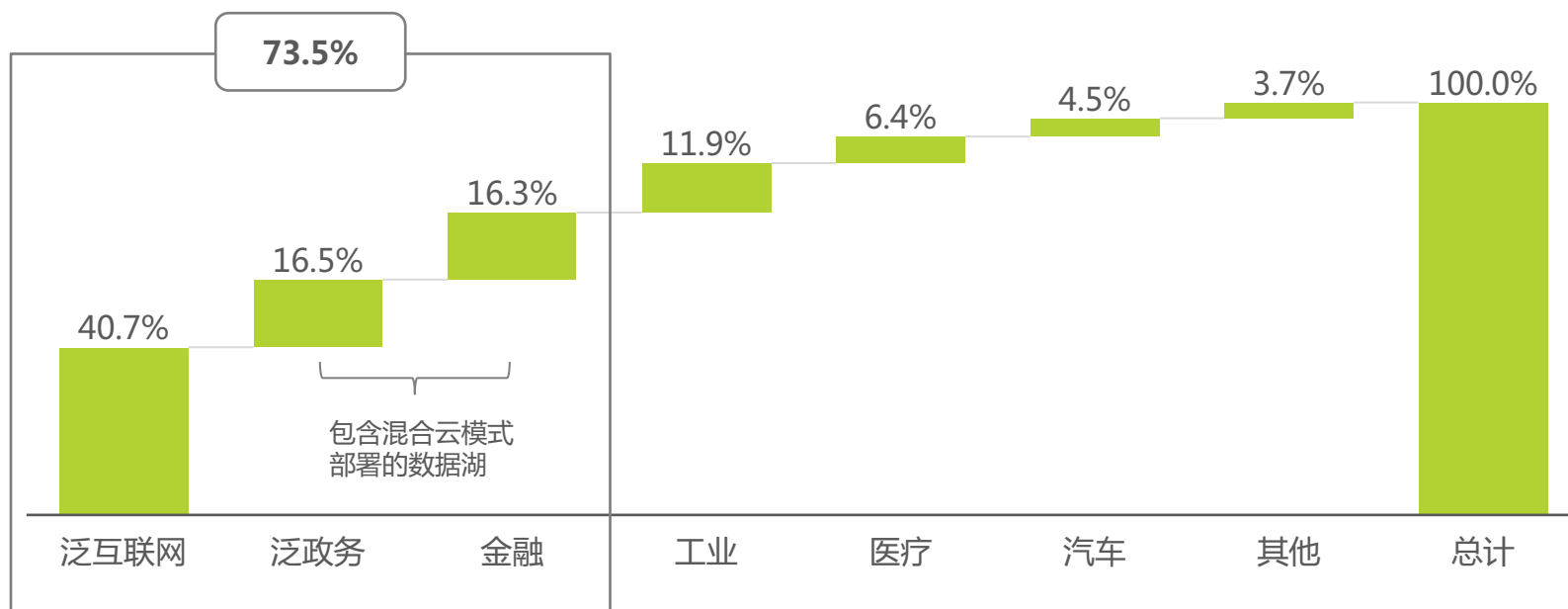
来源：艾瑞咨询研究院自主研究及绘制。

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

现阶段主要应用于泛互联网行业及传统行业的互联网场景

据调研，中国云原生数据湖的下游应用主要分布于泛互联网（电商、网络广告、社交媒体、游戏、互联网金融等）、泛政务（智慧城市、智慧政府、交通等）、金融（银行、保险等）、工业（工业互联网、能源、制造等）、医疗（基因、影像治疗、诊断等）、汽车（车联网等）以及零售、运营商等其他行业。其中，泛互联网企业出于数据量大、非结构化数据多、迭代速度快等原因，率先应用云原生数据湖架构于推荐、搜索、监控等业务环节，是现阶段数据湖市场的主要客户。

2020年中国云原生数据湖市场（不含生态部分）下游行业分布



注释：此下游分布口径为2020自然年全年各厂商在中国内地（不含港澳台）销售云原生数据湖解决方案的营业收入，合同签署地点和交付地点都位于中国内地区域。

注释：此下游分布统计口径不包含云原生数据湖组件发生于其他解决方案（例：数据仓库）的营收，不包含云原生数据湖生态支持厂商部分。

来源：艾瑞长期基础云服务数据监测，结合公开资料、专家访谈，根据数据测算模型，自主研究及绘制。

向更多具有大数据和高价值属性的行业拓展

海量、高频、多源异构的大数据为企业带来了成本、性能和价值挖掘的问题，在现有OLTP数据库+数仓的架构下，企业难以实现底层架构的弹性和优化，无法支持快速发展的业务。云原生数据湖云上部署、存算分离和事后schema的特性可以帮助企业更好地应用数据，未来有望在互联网、汽车、政府、工业等具有大数据和高价值属性的行业得到更广泛的应用。

云原生数据湖的行业应用展望

云原生数据湖解决了什么问题

01 数据海量→成本上升

数据量爆发式的增长，导致对存储和算力资源需求的上升，无论是纵向还是横向扩张，带来成本的叠加都十分惊人。

02 数据多源异构→性能下降

随着互联网的发展，企业外部链接愈发复杂，内部需要处理的数据也愈发多元，包括来自媒体的非结构化数据、web的半结构化数据、物联网的IoT数据、以及来自企业业务系统的结构化数据等。多源异构环境下，数据处理的性能下降，导致企业应用效果不佳。

03 数据价值两极化→实时与聚合

大数据背景下，数据价值愈发向两极聚焦，现有处理架构不能很好地满足实时、聚合分析的诉求，充分释放数据价值。



来源：艾瑞咨询研究院自主研究及绘制。

哪些企业痛点与之匹配

01 业务具有大数据特性，现有架构扩展具有局限性

许多企业在数字化转型的过程中，开始尝试信息流广告、直播电商、远程办公等数字化模式，但底层IT架构和数据架构不能承载海量数据，扩展也存在局限。

02 大数据处理的成本愈发高昂，亟需成本优化解决方案

企业通过增加硬件资源、中间件改造的方式，对数据库、数仓做横向扩展或者纵向优化，成本都十分高昂，企业用于数据的支出日益高企，难以承担。

03 缺乏DT实力和人才，难以进行大数据性能优化，数据价值不能释放

企业缺乏完整的大数据和AI团队，不具备足够的积累去应用前沿大数据和AI技术、搭建面向未来的新架构，故而数据的价值迟迟不能被完全释放，无法实现赋能业务。

未来行业渗透展望

大数据



互联网

互联网企业天然具有大数据的特性，需要云原生数据湖架构来支持业务的快速迭代发展。



汽车

车联网和自动驾驶的数据快速增长，资源扩容与处理速度跟不上业务发展，数据湖应用空间广。

高价值



政府

在政策的驱动下，以智慧城市/政务为中心的信息化建设正在加速推进，城市统一数据中心需求旺盛。



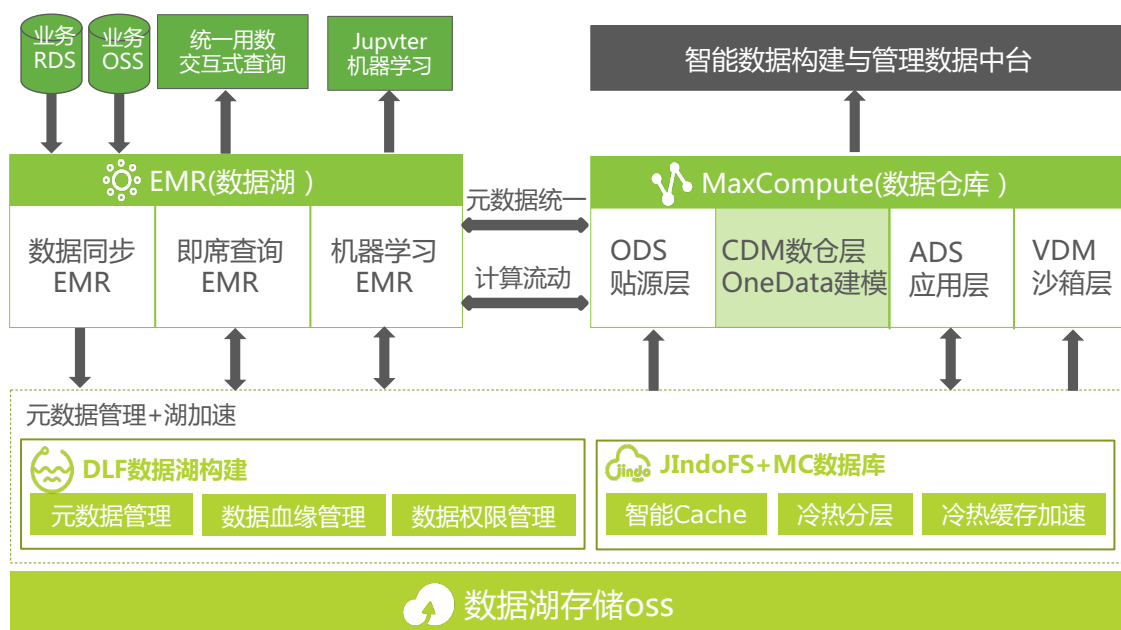
工业

工业数据价值高，标准与治理痛点突出，基于云原生数据湖可以帮助其在云上进行数据统一治理。

计算性能与数据权限隔离能力共同提升，显著降低成本

数禾科技成立于2015年，公司以大数据和技术为驱动，为银行、信托、消费金融公司、保险、小贷公司等持牌金融机构提供高效的智能零售金融解决方案，包括营销获客、风险防控、运营管理等服务，赋能金融机构数字化转型，在消费信贷、小微企业信贷、场景分期、财富管理等多个领域中均有应用。由于金融行业涉及的数据繁多，安全隐私要求程度高，在机构数字化转型过程中，存在运维成本高、数据权限隔离、性能要求高等一系列业务难点。通过与阿里云JindoFS的合作，数禾对数据计算性能的需求得到了满足；同时，围绕Apache Ranger开发权限方案，数禾对数据湖数据权限进行严格管控；利用EMR企业能力协助进行不同部门的资源隔离能力和分账能力；并采用弹性伸缩成本节约模式，兼顾了稳定性和成本。

基于MaxCompute+DLF+EMR+OSS的湖仓一体架构



行业特性&业务难点

- 需要同时运维两套系统，**运维成本过大**；
- 基于HDFS的**存储和不够灵活的计算资源**，成本遇到极大挑战，需要根据任务自动大规模弹性扩缩容；
- 客户服务了大量内部和外部用户，且数据较为敏感，要求严格的**数据权限隔离**；
- 大量OSS的rename等操作，**性能要求高**。

解决方案&产品价值

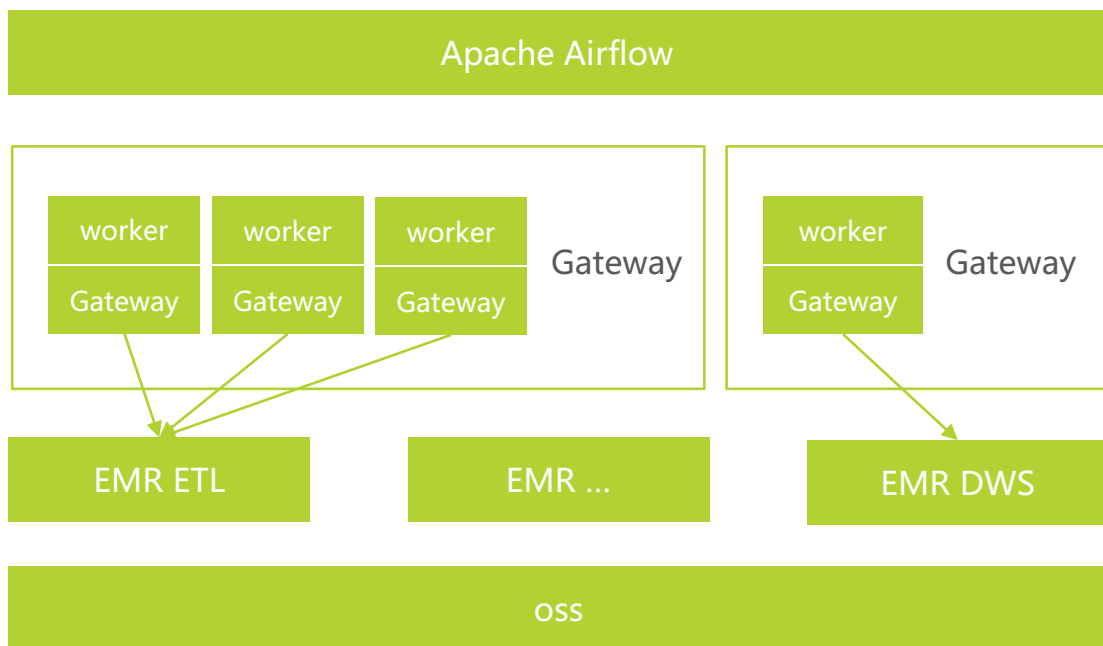
- 通过JindoFS与OSS配合，在存算分离的架构下，满足了用户的数据计算性能需求；
- 围绕**Apache Ranger开发权限方案**，围绕数据湖数据权限严格管控；
- 利用**EMR企业能力如资源组、标签等的支持**，协助不同部门进行资源隔离能力和分账能力的建设；
- 采用弹性伸缩成本节约模式，兼顾稳定性和成本，**压缩成本达20%**。

来源：阿里云，艾瑞咨询研究院自主研究及绘制。

EMR提供计算和存储的弹性拓展能力，助力企业成本优化

流利说成立于2012年9月，是由王翌博士和胡哲人、林晖博士共同创立的科技驱动的教育公司，2018年9月，流利说正式挂牌纽交所，以其独创的教育 3.0 模式，被誉为“AI+教育”第一股。企业希望提高数据质量并完善数据处理方案，提高计算效率。阿里云EMR+OSS云上数据湖架构为企业提供了计算弹性拓展与存储弹性拓展能力，减少了流利说对底层基础设施建设运维的投入。基于阿里云EMR，流利说搭建了Spark、Hive、Presto等大数据处理框架，对存储的数据进行分析，通过智能算法分析学生学习质量，提供相应指导。此外，流利说基于阿里云OSS对在线教育场景下多种类型数据进行集中存储，实现了最大程度的成本优化。

流利说基于EMR+OSS的云上数据湖架构



企业需求

为了提升商业转化效率和公司运营效率，流利说需要打通多业务数据源，统一存储多种应用各类数据。客户受限于数据质量和计算成本，期望借助云计算厂商的能力提高数据质量，优化数据处理方案。

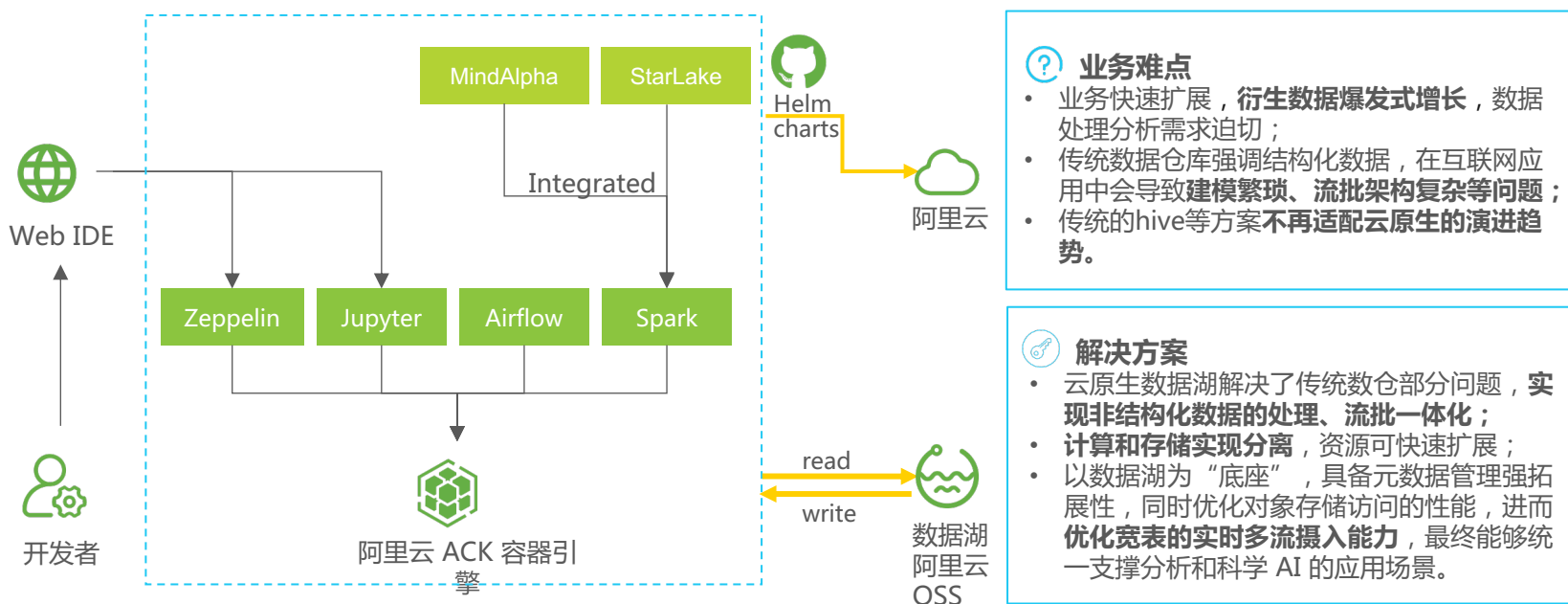
解决方案与效果

- 数据入湖，从DataX全量Dump的方式转变为DataX与Delta相结合的方式，**成本节省70%以上**；
- 数据平台计算集群**成本下降50%**；
- 80%的平台任务从Hive迁移到Spark，**整体任务时间提升30%**。

数据处理能力显著提升，快速构建数据智能应用

汇量科技有限公司（Mobvista）成立于2013年3月，是一个全球性技术平台，通过为企业打造增长赋能的“SaaS工具生态”，提供包括移动营销、统计归因、创意自动化、流量变现、云架构成本优化等一系列产品和服务，助力企业在全球范围内的增长。目前企业数字化转型进入深水区，营销场景往往是转型落地的第一目标。在此过程中，企业业务快速扩展，衍生数据的爆发式增长带来了迫切的数据处理分析需求；而传统数据仓库强调结构化数据，这在互联网应用中会导致建模繁琐、流批架构复杂等问题；同时，传统的hive等方案也不再适配云原生的演进趋势。通过与阿里云数据湖合作，汇量科技实现了非结构化数据的处理、流批一体化；并且使得计算和存储分离，实现了资源的快速扩展；另外，宽表的实时多流摄入能力得以优化，能够统一支撑分析和科学 AI 的应用场景。

汇量科技基于EMR+OSS的云上数据湖架构



业务难点

- 业务快速扩展，衍生数据爆发式增长，数据处理分析需求迫切；
- 传统数据仓库强调结构化数据，在互联网应用中会导致建模繁琐、流批架构复杂等问题；
- 传统的hive等方案不再适配云原生的演进趋势。

解决方案

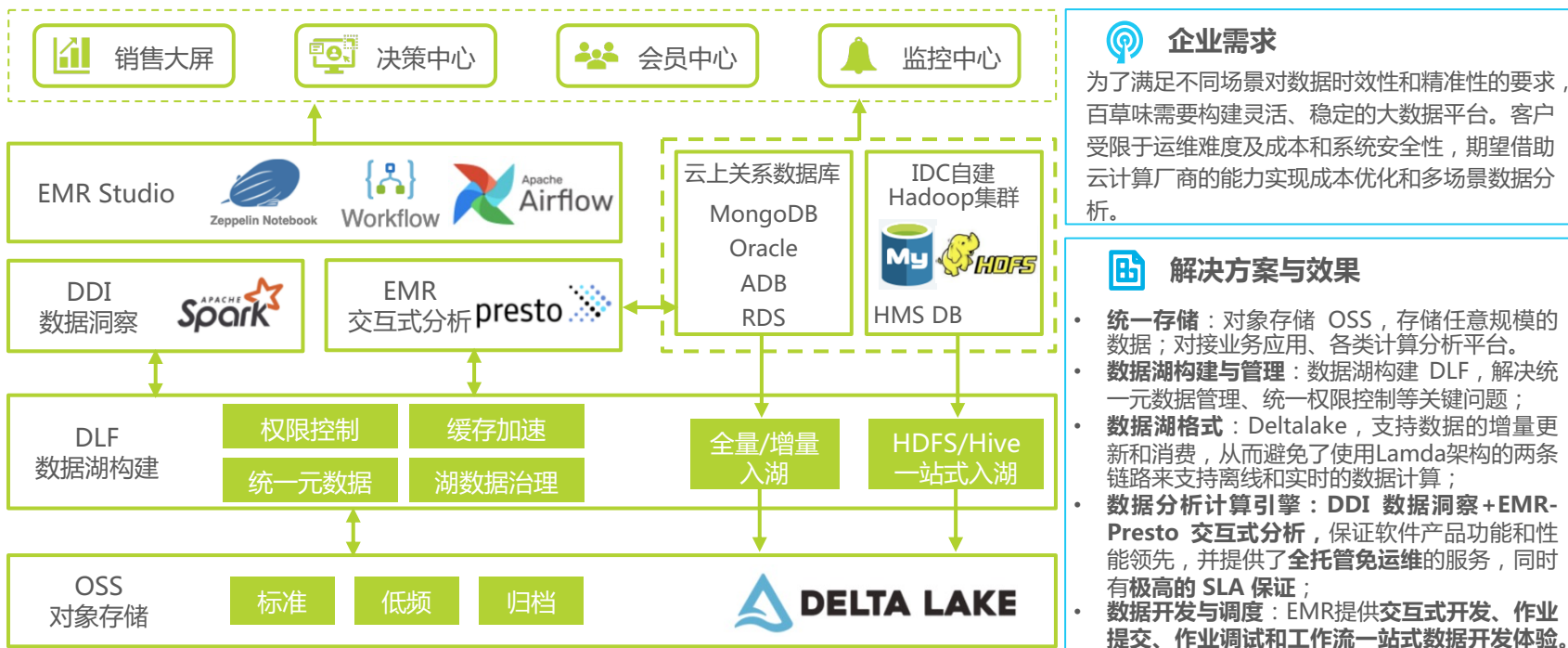
- 云原生数据湖解决了传统数仓部分问题，实现非结构化数据的处理、流批一体化；
- 计算和存储实现分离，资源可快速扩展；
- 以数据湖为“底座”，具备元数据管理强拓展性，同时优化对象存储访问的性能，进而优化宽表的实时多流摄入能力，最终能够统一支撑分析和科学 AI 的应用场景。

来源：阿里云，艾瑞咨询研究院自主研究及绘制。

构建灵活稳定的大数据平台，实时精准响应不同场景需求

百草味是以休闲食品研发、加工、生产、贸易、仓储、物流为主体，集互联网商务经营模式、新零售为一体的全渠道品牌和综合型品牌，目前拥有全品类零食产品1000+SKU，致力于领跑中国休闲食品走向全新格局。企业希望对接多个第三方系统，满足不同场景对数据时效性和精准性的要求，减轻团队工作负担。通过与阿里云的合作，百草味利用对象存储OSS，构建 DLF，实现统一元数据管理和统一权限控制。同时，DDI数据洞察和EMR-Presto交互式分析在保证软件产品功能和性能领先的基础上，还提供了全托管免运维服务，使百草味最终实现实时、精准对接各个场景，全面提高企业运行效率。

百草味基于“EMR+Databricks+DLF”的云上数据湖架构



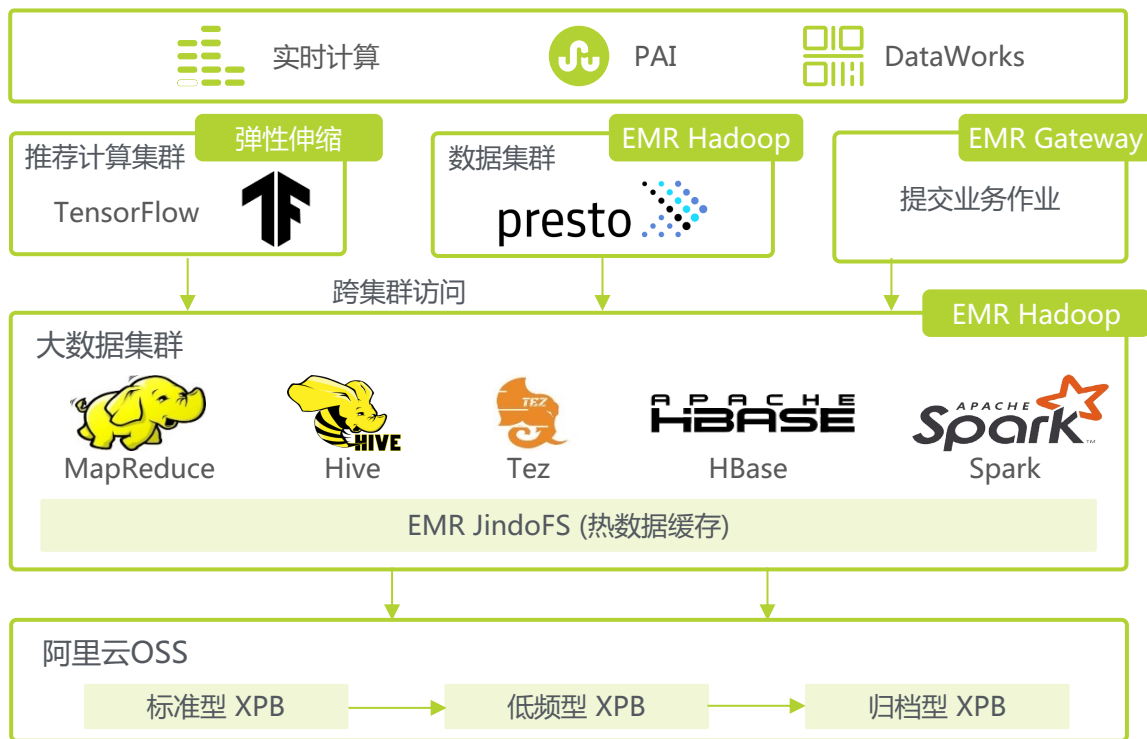
来源：阿里云，艾瑞咨询研究院自主研究及绘制。

互联网社交平台×Soul

提高平台稳定性，降低运维难度，保障APP稳定运营

Soul成立于2016年，是基于兴趣图谱和游戏化玩法的产品设计，属于新一代年轻人的虚拟社交网络，致力于打造一个“年轻人的社交元宇宙”。企业希望提高运维效率，减少ETL任务耗时，建立稳定的系统架构支撑APP在各个时段正常运营。通过与阿里云的合作，Soul利用EMR Delta打造实时数仓，提升了业务指标的实时性；利用JindoFS从HDFS 3副本的架构迁移到OSS，优化了存储成本；同时，通过计存分离降低运维难度和计算成本，最终实现APP的稳定运营。

Soul “EMR+OSS” 的云上数据湖架构



企业需求

为了满足业务高速迭代和业务体量的上涨，Soul需要构建低成本、稳定的平台并降低运维难度。客户受限于人力、工具的短缺和架构的缺失，期望借助云计算厂商的能力，在短期内提升运维效率、优化成本及APP的稳定运营。

解决方案与效果

- 通过EMR Delta打造实时数仓，提升业务指标的实时性，满足更多**实时场景对数据的需求**；
- 利用JindoFS从HDFS 3副本的架构迁移到OSS，以及基于**OSS的归档能力，降低20%的存储成本**；
- 采用**计存分离**的架构，降低计算成本和运维复杂度。

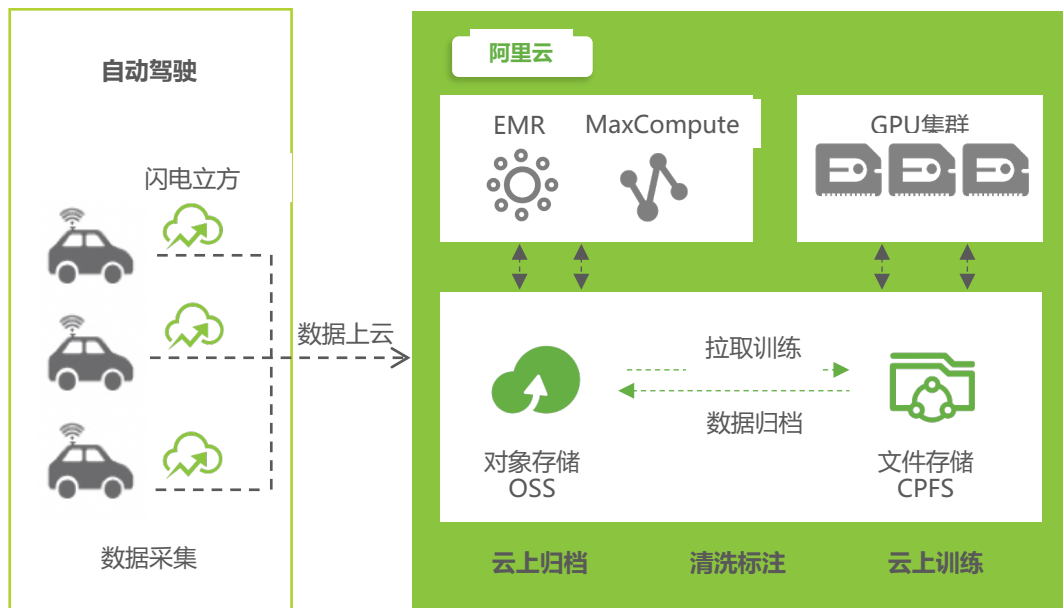
来源：阿里云，艾瑞咨询研究院自主研究及绘制。

自动驾驶×小鹏汽车

数据传输、处理、存储能力同步提升，轻松实现各种训练

小鹏汽车正式成立于2015年，是一家专注未来出行的科技公司，目前已成为中国领先的智能电动汽车公司之一。公司一直坚持饱和式研发投入，构建全栈自研的核心能力，致力于用科技为人类创造更便捷愉悦的出行生活。在智能化、网联化、电动化、共享化的背景下，自动驾驶成为智能网联汽车行业的重点，也成为了下一代汽车行业转型升级的技术高地。自动驾驶过程中，车辆每天会产生大量采集数据。对于这些分布在不同地域的数据，如何及时合规地存储以及高效便捷地计算是一大业务难点。通过与阿里云闪电立方合作，多区域上传节点，小鹏汽车实现了大批量采集数据快速上传到云上数据湖；进入湖中的采集数据，通过云上EMR、Maxcompute进行大规模处理和标注，处理后的数据持久存储到OSS；数据湖通过与文件存储CPFS数据流动，湖中数据更加轻松的与GPU算力对接，实现各种训练。

小鹏汽车云上数据湖架构



业务难点

- 车辆每天产生大量采集数据，这些分布在不同的地域数据需要及时完成合规存储，让采集设备能投入下一个采集周期；
- 存储的数据需要能有便捷的方式与计算能力对接，应用到自动驾驶数据清洗、标注、训练等多种不同场景中；
- 需要丰富的计算引擎和强大算力来全面覆盖仿真、训练、标注等各种数据处理与分析场景。

解决方案

- 阿里云闪电立方解决了自动驾驶车辆终端采集难题，通过阿里云多区域上传节点，大批量采集数据得以快速上传到云上数据湖；
- 进入到数据湖的采集数据，通过云上EMR、Maxcompute进行大规模的数据处理和标注，处理后的数据持久存储到OSS；
- 数据湖通过与文件存储CPFS数据流动，让数据湖中数据更加轻松地与GPU算力对接，实现各种训练，训练后的数据再归档到OSS，高性能文件存储只需要存储临时少量热数据。

来源：艾瑞咨询研究院自主研究及绘制。

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

建议一：战略规划

建立统一的数据底座，支持企业向数据驱动转型

对于现代化企业来说，需要面对愈发复杂多元、高频迭代的内外部环境，仅依靠人力难以跟上市场的发展，“数据驱动”成为企业的必然选择。而“数据驱动”落在实践中还存在很多的问题，并非根据现在的业务需求，采购一些数字化工具即可完成的转型。针对具有“变化、挖掘、未知”特性的需求，企业需要建立统一、弹性、智能的数据底座，以“不变应万变”，从而支持数据驱动，让数据释放价值。

建立基于统一底座的数据驱动策略

现代化企业面临来自内外部的挑战

外部竞争：

现代化企业面临越发易变、模糊、不确定、复杂的外部环境。这从外部驱动企业业务和应用也必须快速迭代，及时响应客户，才能在快速发展的市场上获得优势。

内部管理：

随着企业的发展和多轮信息化改造、数字化升级，内部IT部署很难保持一致和清晰。无论是部署环境，数据存算，还是业务应用都在某种程度上呈现“混乱”的状态，造成了降本增效的困难。

技术部署：

为了应对越来越多种类的数据，以及越来越复杂场景的诉求，大数据、AI技术栈呈指数增长。多种框架并存是未来IT的必然状态，为企业带来了技术部署的挑战。

发展创新：

除了基于现有IT资源和业务进行经营性的“降本提效”外，现代化企业还需要考虑差异化竞争力的打造，通过技术、产品、商业创新，发展第二增长曲线。而通过数据驱动寻找创新点，在为企业带来机遇的同时也提出了更高的要求。

统一数据底座对“数据驱动”的重要性



建议二：执行路线

站在长期视角，着重考虑DT能力的开放性、敏捷性与创新性

在市场快速发展的背景下，企业进行DT能力建设时，需要更加看重技术路线的开放性和扩展性，为难以预测的未来探索做好准备，去支持应用和业务的创新。在应用实施及之后的运营时，企业开发者一方面可以更开放的态度去拥抱云原生与开源，另一方面可以对自身的技术进行抽象、分层和服务化，以更简单的方式提升效率和效益。云原生数据湖架构开放、敏捷，是企业建设DT能力很好的选择之一。

企业DT能力建设的执行实践

➤ 如何保持DT能力的敏捷与创新？



拥抱云原生

云原生是后云计算时代新一轮生产力的释放，包含容器、微服务、Serverless、DevOps等天然具有敏捷弹性优势的技术、工具和方法论，是IT发展的必然趋势。基于云原生，进行数据存储、计算、治理、架构等方面的优化和创新，是大数据发展的必然之路。



拥抱开源

开源是学习成本缩减、技术创新加速的高效生产方式，愈发被国内市场所接受。开发者可以通过开源社区快速、低成本的学习前沿技术。对于缺乏IT积累和专业人才的企业，可以选择更开源兼容的服务商共同进步。



抽象、分层与服务化

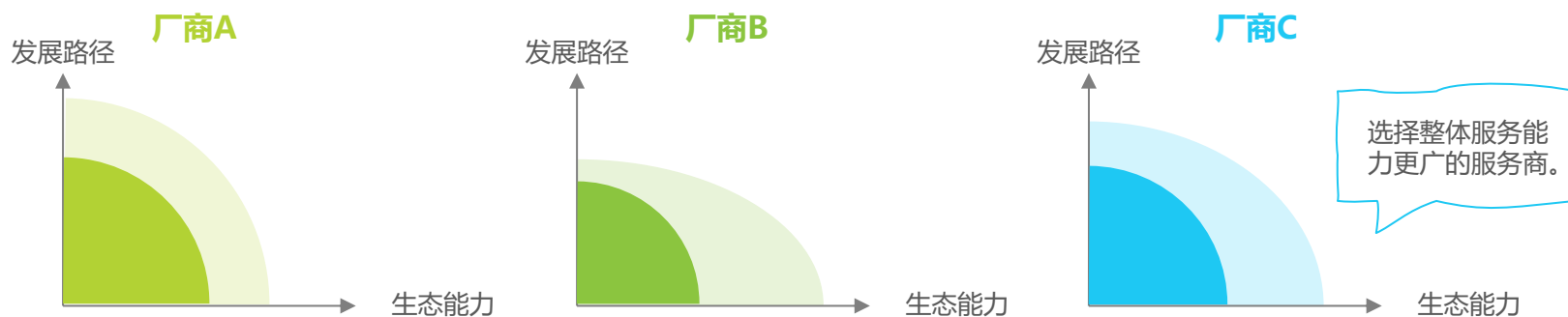
站在开发者视角，需要意识到企业应用和大数据的标准在短时间内是难以实现统一的，且很难回到过去一家之言成为行业标准时代。为了更好地应对标准和需求的复杂，企业可以对自己的IT能力进行抽象和分层，通过模块的标准化来实现效率，通过组合的创新来实现效益，通过交付服务化实现便捷。

建议三：具体选型

选择服务半径更广，发展路径更契合的服务商

云原生数据湖是企业级的综合大数据解决方案，且实践具有长期性，伴随企业的长期IT能力升级。故而除了内部能力（技术、产品、解决方案等）的评估外，云原生数据湖选型还需要格外关注厂商的外部能力和未来能力：是否有足够丰富的生态合作伙伴来满足企业不同场景的需求？技术演进路线是否与企业匹配？是否能支持企业业务未来的拓展？企业需要更综合的考虑，选择整体服务能力更广的服务商。

云原生数据湖的选型矩阵



选型矩阵阐述

1. 内部能力评估

评估厂商本身的能力，包括云原生数据湖核心技术组件（存储、计算、管理等）的性能和功能，以及整体解决方案的成熟性和性价比。

2. 生态能力评估

云原生数据湖不是单一的存储或者数据库产品，而是面向企业大数据应用的全生命周期解决方案。故而，企业在进行选型时，除了厂商本身的能力，还需要关注厂商的生态能力，是否有足够的生态合作伙伴来共同支持企业的多元需求。

3. 未来能力评估

云原生数据湖的部署并非一次性结束的短期项目，涉及企业长期数据能力的发展，故而企业在选型时还需要关注厂商未来的发展路径是否与自己的发展路线契合，是否能支持自身业务未来的拓展。

来源：艾瑞咨询研究院自主研究及绘制。

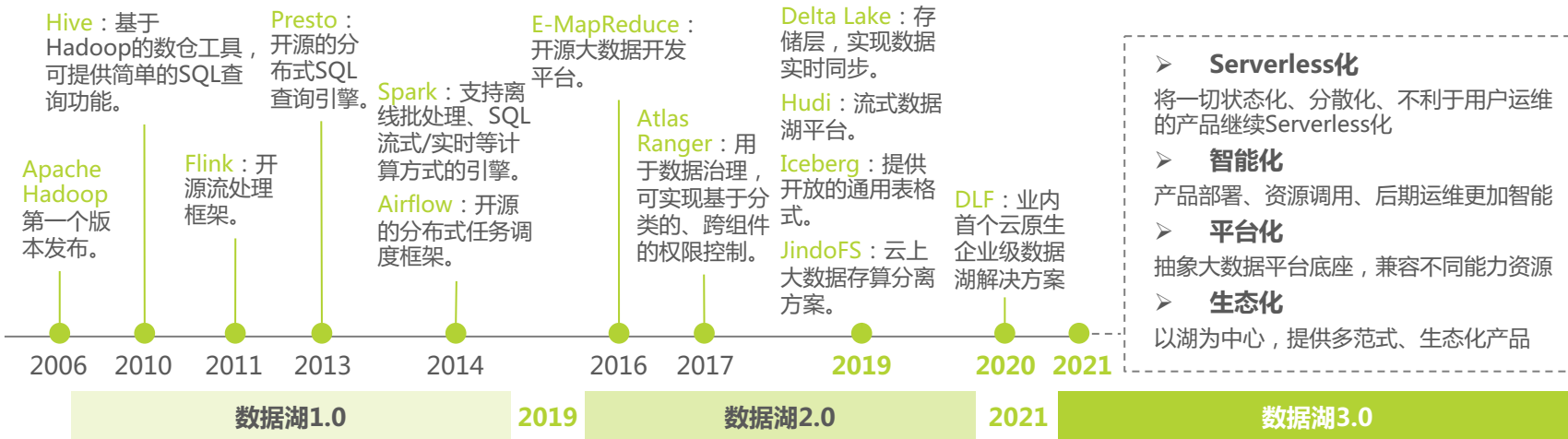
典型企业展示

- 阿里云
- Databricks
- Cloudera

率先入局数据湖市场，持续迭代服务全行业客户数字化升级

面对企业数据治理挑战严峻、产品部署成本高企、大数据管理实施复杂和落地效果不及预期的痛点，作为中国云计算与大数据前瞻的践行者，阿里云率先布局数据湖领域。基于十余年大数据技术的实践与探索，阿里云云原生数据湖解决方案不断迭代升级，至今已经历了三代发展，实现了存储服务化、管理智能化、计算多元化等方面的日益完善，具备松耦合、积木化、广兼容及低运维的优势。在演进的过程中，阿里云积累了互娱、社区、电商、金融、制造等多行业的服务案例，未来将在serverless化、智能化、实时化、平台化、生态化等方面继续深耕，持续赋能全行业客户的数字化转型升级。

阿里云云原生数据湖解决方案：发展历程与演进路线



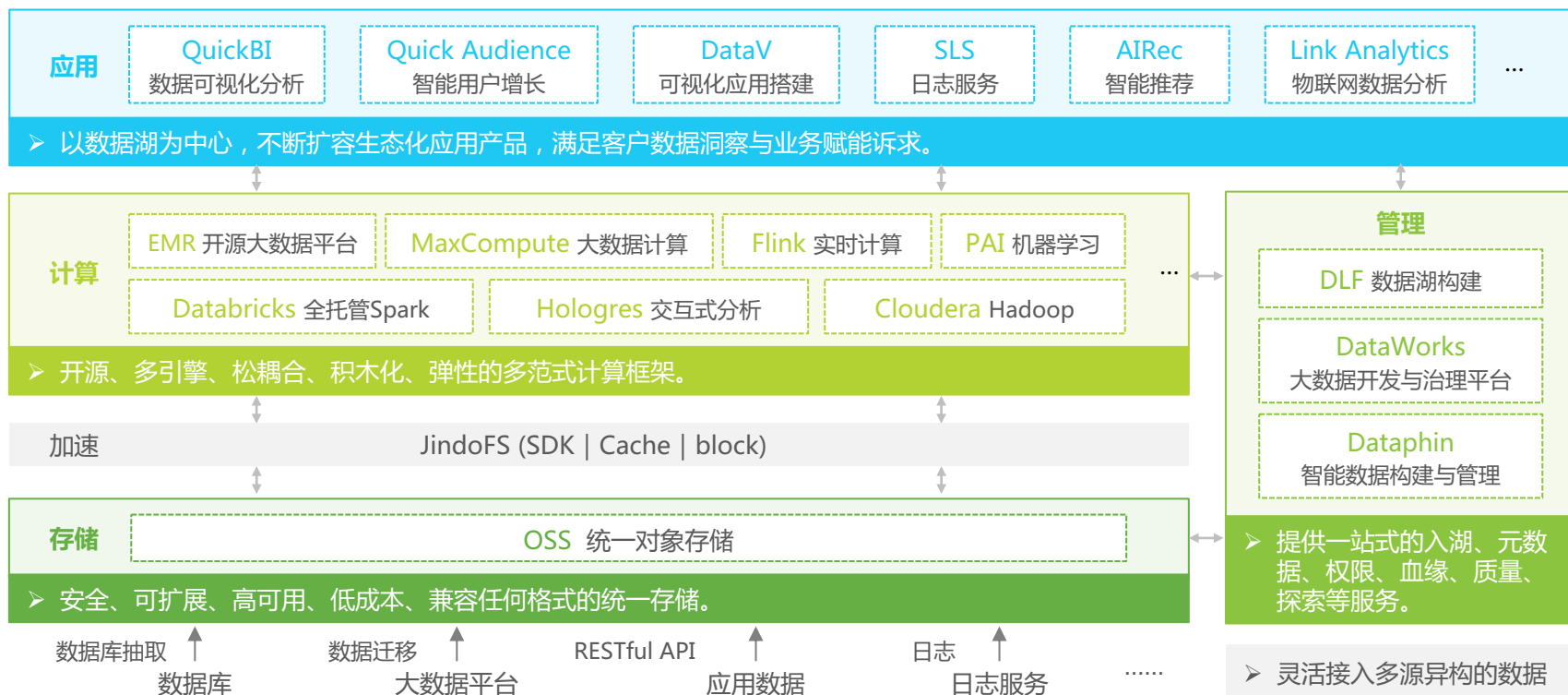
	2006 - 2019 数据湖1.0	2019 - 2021 数据湖2.0	2021 数据湖3.0
存储	存算分离，冷热分层，以Hadoop生态为主。	对象存储为中心，统一存储承载生产业务，大规模、高性能。	对象存储OSS为中心，构建企业级数据湖，全兼容、多协议、统一元数据。
计算	引擎以Hadoop和Spark为主，初步实现云原生，但缺乏弹性及多样性。	云原生、弹性化，但用户仍需在计算侧进行自行搭建，且数据无法统一管理。	与DLF、EMR等计算引擎无缝对接，不仅云原生、弹性化，同时实时化、AI化、生态化。
管理	热数据存储的Hadoop需要投入大量管理硬件运维和扩容任务。	元数据管理和协议转换需用户自行搭建，数据管理无法和对象存储实现无缝融合。	智能“建湖”和“治湖”：面向湖存储+计算的一站式湖构建和管理。

来源：阿里云，艾瑞咨询研究院自主研究及绘制。

兼容、弹性、一站式的大数据架构，满足企业多元洞察诉求

基于云原生相关技术和计算存储分离架构，阿里云推出了云原生企业级数据湖解决方案。在该架构下，数据湖**直接对接**企业业务生产中心多源异构的海量数据，**统一存储**于阿里云对象存储OSS，**弹性调用**阿里云EMR、MaxCompute、PAI，以及Flink、Spark等主流开源计算引擎，**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。

阿里云云原生数据湖解决方案：架构与优势

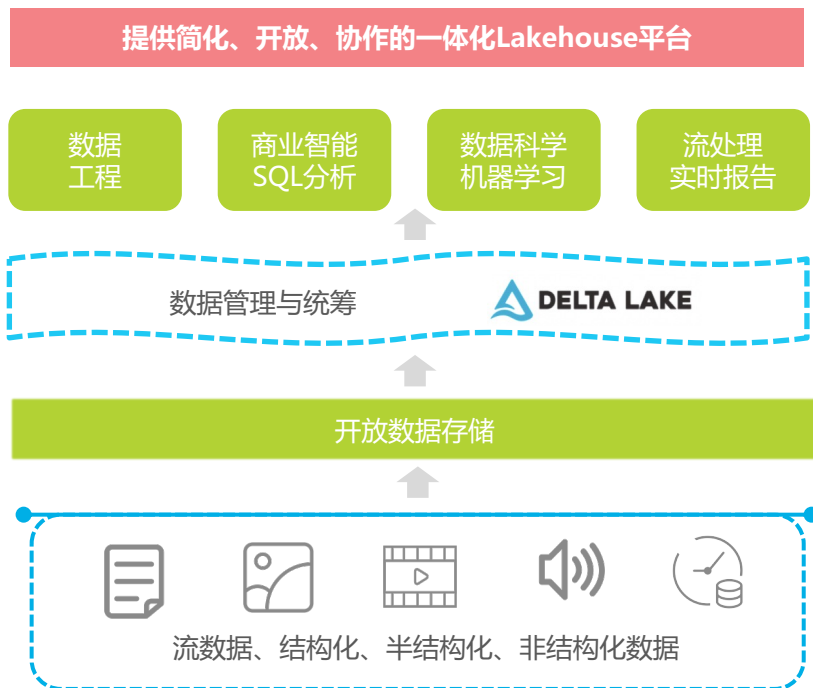


来源：阿里云，艾瑞咨询研究院自主研究及绘制。

湖仓一体架构，统一管理所有数据、分析和AI工作

Databricks旗下的Lakehouse平台基于湖仓一体架构，实现了数据湖和数据仓库最佳实践的结合。Lakehouse平台由全新的开放和标准化系统设计支持：直接在用于数据湖的低成本存储类型上，同时实现数据仓库中的数据结构和数据管理功能。这个统一的平台消除了传统方案中，将分析、数据科学和机器学习分开的数据孤岛困境，进一步简化了数据架构。它建立在开源和开放标准体系之上，以最大限度地提高灵活性。另外lakehouse凭借其原生协作功能可以提高跨团队工作的效率，并触发更多的创新。

Databricks Lakehouse架构及优势



- 统一部署 扩展性强**

在一个通用平台上**统一所有的数据、分析和 AI 工作项目**来简化数据架构。Delta Lake提供支持所有类型数据以及数据分析工作所需的可靠性、性能和治理，构成了这个具有成本效益、高度可扩展的湖仓一体的基础。
- 开放标准 多方兼容**

该平台建立在世界上最成功的开源数据项目之上，将数据生态系统与开放标准和格式统一起来。这意味着可以**避免专有数据格式**，并**无障碍使用**合作伙伴提供的软件和生态系统。
- 原生协作 跨域支持**

借助原生协作功能，可以**统一数据团队**以在整个“数据+AI”工作流中进行协作。从数据工程师到分析师再到数据科学家，都可以**跨工作任务进行协作**，从而加快数据驱动进程。
- 实时报告 端到端流**

实时报告是许多企业的常态需求。**对流处理的支持**消除了对那些专门服务实时数据应用程序系统的需要，简化平台系统负载。
- BI支持 存算分离**

支持**直接在源数据上使用 BI 工具**，减少了延迟；存储与计算分离，这意味着存储和计算使用单独的集群，因此这些系统能够**扩展到更多并发用户和更大的数据容量**。

来源：艾瑞咨询研究院自主研究及绘制。

加速“数据+AI”进程，降低项目风险，快速取得成功

Databricks旗下的Lakehouse一体化平台功能齐全，可以提供跨行业和跨角色的数据分析和人工智能解决方案，解决客户面临的最常见问题以及一些有重大影响的应用难题。其用例的应用场景十分广阔，比如在商业领域追踪供应链全流程数据进行库存调整和更快、更准确的需求预测，以及在医疗保健方面进行数字病理图像分析和疾病传播预测等。Databricks帮助数据科学家们开始“数据+AI”的协作转型，大规模构建快速可靠的数据管道。与Hadoop架构相比，Databricks大幅降低了使用成本，增加了数据团队的有效产出，进而利用数据驱动创新助力了客户成功。

Databricks Lakehouse行业应用



来源：艾瑞咨询研究院自主研究及绘制。

CDP顺应大数据架构融合的趋势，为企业提供存算分离、流批一体的云原生解决方案

Cloudera成立于2008年，是第一批提供企业级统一大数据平台的厂商之一。2018年，Cloudera 在与 HortonWorks 合并后，推出了新一代大数据平台Cloudera Data Platform (CDP)。CDP支持混合云和多云环境部署，采用存算分离架构，流批一体计算框架，为企业提供一致的数据分析体验，从而实现现有数据投资价值的扩展和数据洞察力的提升。

Cloudera Data Platform (CDP) 产品架构



来源：艾瑞咨询研究院自主研究及绘制。

企业级数据云，加速金融、运营商等各行业数字化转型

Cloudera致力于帮助各行各业的企业打造DT能力，从不断增长的数据量中获得准确的实时洞察，从而于激烈的市场竞争中脱颖而出。公司服务行业涉及金融服务、电信行业、公共部门、医疗保健、技术、制造业等，其中包括梅哈里医学院、荷兰合作银行、新加坡大华银行等知名企业。

Cloudera的行业应用（部分）

金融	运营商	政府	公安	零售	制造	能源	医疗
<ul style="list-style-type: none"> • 欺诈探测 • 反洗钱 • 风险管理 • 保险定价 • 实时监控 	<ul style="list-style-type: none"> • 经营分析 • 网络监控与分析 • 网络优化 • 流失率分析 • 业务优化 	<ul style="list-style-type: none"> • 法规实施 • 交通流量优化 • 舆情分析 • 雾霾天气预测 	<ul style="list-style-type: none"> • 人脸识别 • 伪车牌识别 • 轨迹分析 • 重点人员管控 • 嫌疑人员挖掘 • 预警 	<ul style="list-style-type: none"> • 精准营销平台 • 动态定价 • 会员数据挖掘 • 门店优化选址 • 全渠道分析 • 销售预测 	<ul style="list-style-type: none"> • 良率分析 • 次品跟踪 • 供应链优化 • RFID关联分析 • 授权管理 • 主动性维护 	<ul style="list-style-type: none"> • 自然资源探测 • 加油站画像 • 库存优化 	<ul style="list-style-type: none"> • 药物开发 • 科学研究 • 临床研究 • 病例共享 • 健康结果分析
 <h3>跨行业应用场景</h3>							
<ul style="list-style-type: none"> • 情感分析 • 社交CRM/网络分析 • 流失率缓解 • 品牌监控 • 跨界与销售提升 • 忠诚度与促销分析 • Web应用优化 		<ul style="list-style-type: none"> • 市场活动优化 • 品牌管理 • 社交媒体优化 • 价格优化 • 内部风险评估 • 营业额担保 			<ul style="list-style-type: none"> • 物流优化 • 点击流分析 • 影响力分析 • IT架构分析 • 法律发现 • 设备监控 • 企业搜索 		

来源：艾瑞咨询研究院自主研究及绘制。

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

云原生与大数据背景下，数据湖成为企业智胜未来的新一代生产力工具，市场将迎来爆发期

在数据生产或处理爆炸性增长、实时化、智能化、云化等背景下，数据湖凭借“兼容、松耦合、弹性、敏捷”的天然优势赋能云原生时代的大数据治理，俨然成为了新的“掘金热土”，未来在大数据产业中的占比将持续上升，前景广阔。

云原生时代的大数据治理——数据湖

数据生产和处理正在发生质变

01 规模爆炸性增长

40ZB

2020年全球数据规模

430%

2020-2025年
全球数据规模增长

02 生产/处理实时化

30%

2025年实时数据占比

50%

2022年新业务将会采用
实时分析比例

03 生产/处理智能化

80%

2025年全球
非结构化数据占比

55%

全球非结构化数据增速

04 数据加速上云

32.7%

2020年中国
公有云数据库市场份额

47.2%

2025年中国公有云
数据库市场规模

数据湖天然适合云原生时代的大数据治理

数据

企业数据量高速增长，半结构化和非结构化占比增长，数据价值不断被挖掘。

存储

原始文件的集中存储和调用，有助于：

- 降低成本、简化运维；
- 企业数据的天然打通

计算

不管是计算引擎本身，还是人工智能算法，都趋向于多元，以适应企业不同场景。

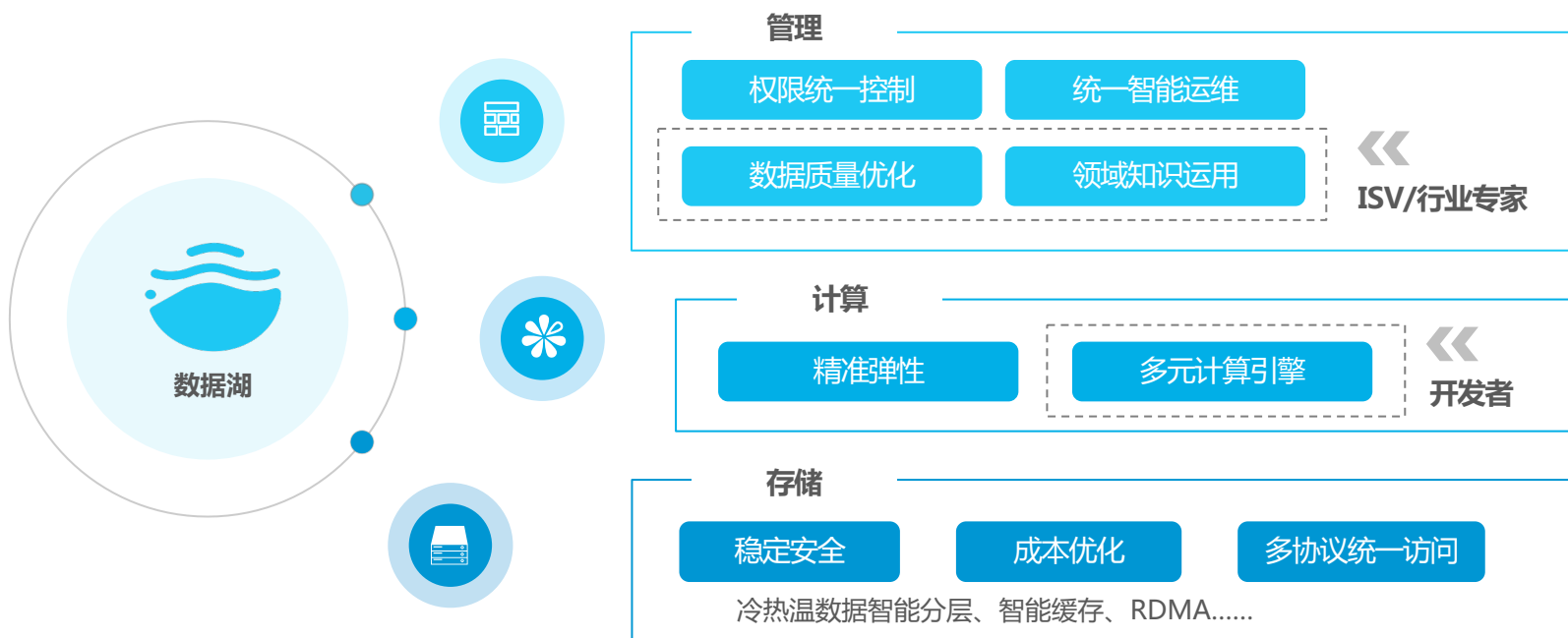
企业敏捷性创新

- ① 读时模式（Schema后置）有助于按需即取；
- ② 计算与存储的松耦合性，对开源产品的拥抱性，有助于：
 - 企业采用混合架构及平滑上云；
 - 降低企业IT人才风险

生态共赢，驱动云原生数据湖解决方案日臻完善

尽管数据湖与云和大数据天然契合（海量、弹性、简单、敏捷），但在企业业务场景落地中，仍有许多实际问题需要解决。例如，弹性计算与实际业务量的精确拟合，数据权限的精细控制，冷热温数据的精准分层，数据质量的提升、领域知识的建立等等。这些复杂而琐碎的问题并非一个厂商单独可以解决，需要在数据湖厂商、开发者、ISV和SI的共同努力，在企业级生产环境中不断探索与摸索，走向完善与繁荣。

云原生数据湖基础架构演进过程中的要求



艾瑞新经济产业研究解决方案



行业咨询

- 市场进入 为企业提供市场进入机会扫描，可行性分析及路径规划
- 竞争策略 为企业提供竞争策略制定，帮助企业构建长期竞争壁垒



投资研究

- IPO行业顾问 为企业提供上市招股书编撰及相关工作流程中的行业顾问服务
- 募 投 为企业提供融资、上市中的募投报告撰写及咨询服务
- 商业尽职调查 为投资机构提供拟投标的所在行业的基本面研究、标的项目的机会收益风险等方面的深度调查
- 投后战略咨询 为投资机构提供投后项目的跟踪评估，包括盈利能力、风险情况、行业竞对表现、未来战略等方向。协助投资机构为投后项目公司的长期经营增长提供咨询服务

关于艾瑞



艾瑞咨询是中国新经济与产业数字化洞察研究咨询服务领域的领导品牌，为客户提供专业的行业分析、数据洞察、市场研究、战略咨询及数字化解决方案，助力客户提升认知水平、盈利能力和综合竞争力。

自2002年成立至今，累计发布超过3000份行业研究报告，在互联网、新经济领域的研究覆盖能力处于行业领先水平。

如今，艾瑞咨询一直致力于通过科技与数据手段，并结合外部数据、客户反馈数据、内部运营数据等全域数据的收集与分析，提升客户的商业决策效率。并通过系统的数字产业、产业数据化研究及全面的供应商选择，帮助客户制定数字化战略以及落地数字化解决方案，提升客户运营效率。

未来，艾瑞咨询将持续深耕商业决策服务领域，致力于成为解决商业决策问题的顶级服务机构。

联系我们 Contact Us

 400 - 026 - 2099

 ask@iresearch.com.cn



企 业 微 信



微 信 公 众 号

法律声明

版权声明

本报告为艾瑞咨询制作，其版权归属艾瑞咨询，没有经过艾瑞咨询的书面许可，任何组织和个人不得以任何形式复制、传播或输出中华人民共和国境外。任何未经授权使用本报告的相关商业行为都将违反《中华人民共和国著作权法》和其他法律法规以及有关国际公约的规定。

免责条款

本报告中行业数据及相关市场预测主要为公司研究员采用桌面研究、行业访谈、市场调查及其他研究方法，部分文字和数据采集于公开信息，并且结合艾瑞监测产品数据，通过艾瑞统计预测模型估算获得；企业数据主要为访谈获得，艾瑞咨询对该等信息的准确性、完整性或可靠性作尽最大努力的追求，但不作任何保证。在任何情况下，本报告中的信息或所表述的观点均不构成任何建议。

本报告中发布的调研数据采用样本调研方法，其数据结果受到样本的影响。由于调研方法及样本的限制，调查资料收集范围的限制，该数据仅代表调研时间和人群的基本状况，仅服务于当前的调研目的，为市场和客户提供基本参考。受研究方法和数据获取资源的限制，本报告只提供给用户作为市场参考资料，本公司对该报告的数据和观点不承担法律责任。

为商业决策赋能

EMPOWER BUSINESS DECISIONS

