中国数智融合发展洞察

摘要

VUCA时代，市场变化加速。企业需要更加敏捷而准确的数智化决策，这些决策应当是分钟级的而非天级的，应当是基于全量数据的而非局部数据的，应当是基于准确数据的而非基于“脏数据”的，应当是业务人员和数据分析人员任意发起的而非是通过复杂流程和多部门配合才能实现的。

传统的数仓或者湖仓分离架构让数智融合和企业敏捷决策变得困难：数据孤岛存在，决策无法基于全量数据；数据来回流转，成本高、周期长、时效差。基于存储-缓存-计算分离，湖-仓-AI数据统一元数据管理的Serverless，可在数据量、成本、效率、敏捷方面取得最优解。

开源为数智生态贡献重要力量，但这不预示所有企业需通过开源产品自建数智平台。实际上，大多企业聚焦自己核心业务，选择性能稳定、无须运维、数智融合、端到端自动化与智能化的商业化数智平台，ROI会更高。当然，平台应与主流开源产品具有良好继承性，如此，更加灵活开放，企业的IT人才补给成本也更低。

来源：艾瑞咨询研究院自主研究及绘制。

中国数智融合发展背景 1

企业数智融合的痛点及应对 2

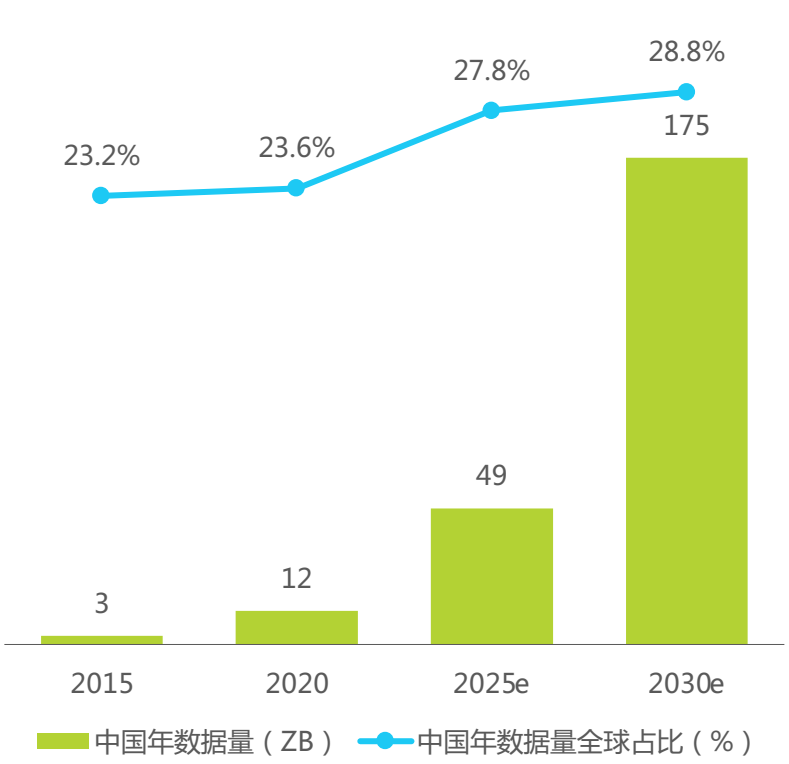
数智融合典型实践 3

数据量和非结构化数据占比上升

统一管理，统一查询使用，成为新的挑战

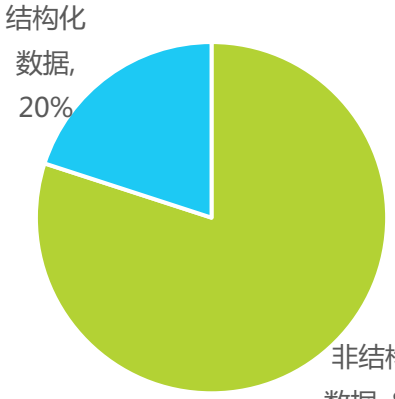
全球数据量以59%以上的年增长率快速增长，其中80%是非结构化和半结构化数据，中国数据量的上升较全球更为迅速。数据量和非结构化数据的上升，使得基于对象存储的数据湖越来越为普及。此时，如何使用统一管理，统一查询使用，成为新的挑战。

2015-2030年中国数据量规模及全球占比



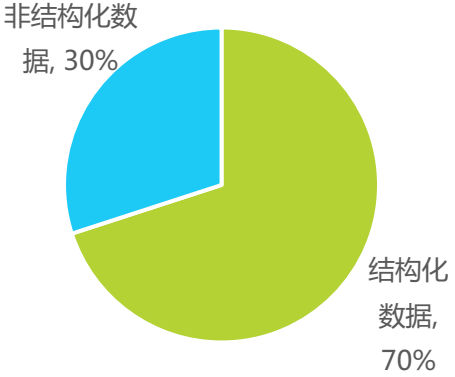
来源：中国电信招股说明书，艾瑞咨询研究院整理及绘制。

企业内结构化数据与非结构化数据占比及使用情况



在企业的数据中，结构化数据仅占20%，其余80%都是以文件、语音、图片等形式存在的非结构化数据。且非结构化数据的增速远远高于结构化数据，随着时间的推移,非结构化数据所占的比例将会越来越高。

企业长期以来，受技术影响，对结构化数据的利用率均高于对非结构化数据的利用率。但实际上，非结构化数据的体量与其包含的信息量都更多，是企业未得到充分利用的宝贵资产。



来源：艾瑞咨询研究院自主研究及绘制。

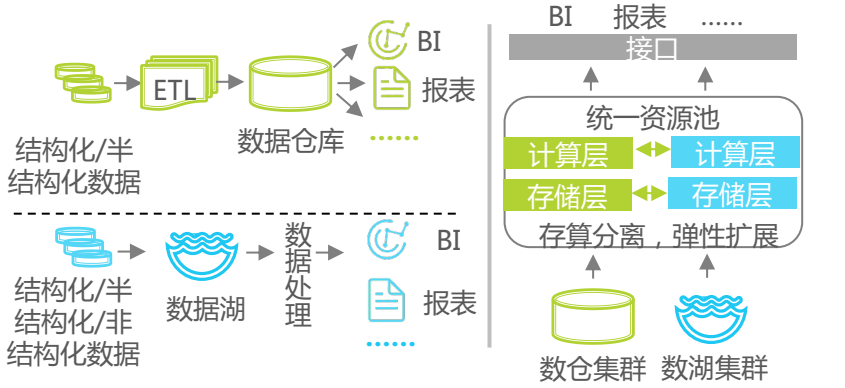
数据多源异构成为常态

数据从“汇聚才可被用”到“链接即可被用”

在传统数仓中，多源数据经ETL过程并集中入仓，方可被使用。该方式有许多不足：第一，因有复杂的ETL过程及大量数据的传输，数据实时性难以保障，因此分析常必须T+1才可完成；第二，数据的全量存储和存储成本之间难以取舍，因此必须提前抉择保留哪些数据，随着数据种类的逐渐增多，这很难做到；第三，对于异常值的下钻、回溯等，无法回溯到最为原始的数据。随着应用场景的增多，数据库的种类也逐渐丰富，如更适应物联网场景的时序数据库、更适应知识谱图应用的图数据库，等等。

综上，多源异构、分布存储、现用现传、统一查询与应用的架构，逐渐被敏捷型企业认可。

数据仓库vs数据湖vs湖仓一体



数据仓库vs湖仓一体

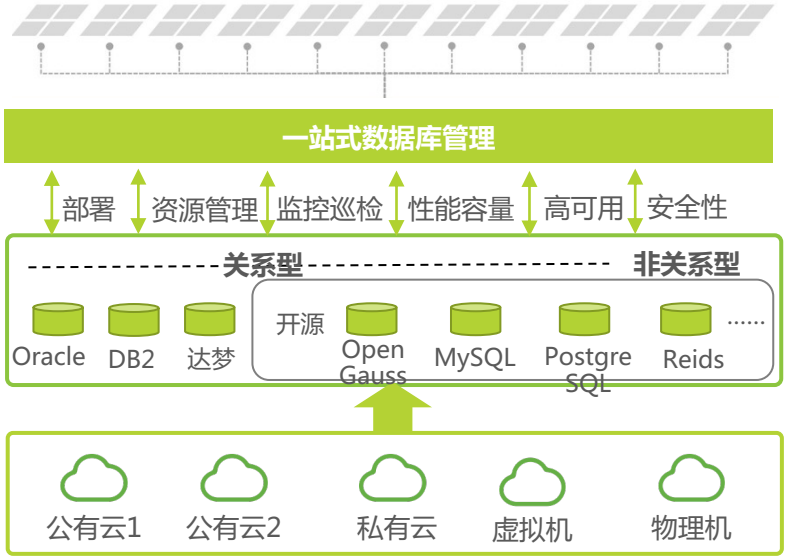
“湖仓一体”作为数据处理统一底座，提供实时处理多引擎、多数据类型能力，避免数据移动建模，降低数据处理的成本。

数据湖vs湖仓一体

“湖仓一体”弥补Hadoop下数据湖实时数据处理的缺失，降低事后数据治理难度，提升了大数据应用性能。

来源：艾瑞咨询研究院自主研究及绘制。

数据库的多源性



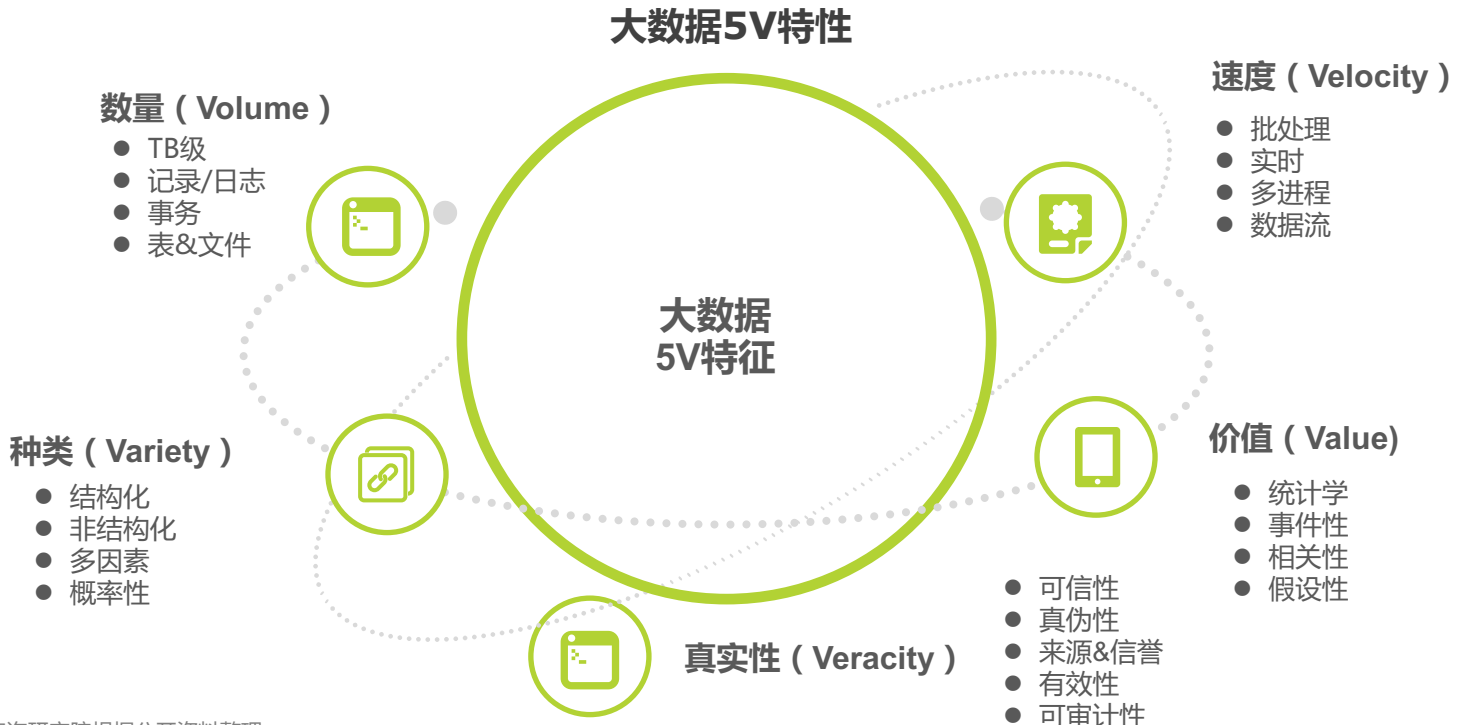
来源：艾瑞咨询研究院自主研究及绘制。

©2022.7 iResearch Inc. www.iresearch.com.cn

大数据的5V价值有待进一步释放

可从平台性工具入手，进而解决思维和技能的问题

大数据产业作为以数据生成、采集、存储、加工、分析、服务为主的战略性新兴产业，提供全链条技术、工具和平台，孕育数据要素市场主体，深度参与数据要素全生命周期活动，是激活数据要素潜能的关键支撑，是数据要素市场培育的重要内容。目前，大数据产业仍存在数据壁垒突出、碎片化问题严重等瓶颈约束，大数据容量大、类型多、速度快、精度准、价值高的5V特性未能得到充分释放。这其中既有思维、技能的要素，又有工具的要素，三者也并非割裂存在，一般来说，性能稳定、简单易用的全链条平台工具有助于消除思维的“不敢”和技能的“不会”，化解掉5V特性释放的原始阻力，使得大数据更加普适化。



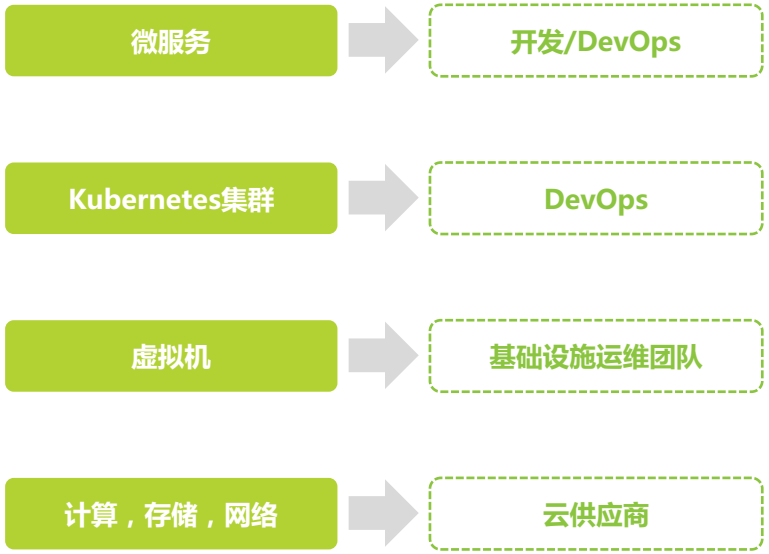
来源：艾瑞咨询研究院根据公开资料整理。

云原生：从微服务走向Serverless

从PaaS到FaaS，基础设施被更深层次地托管和“屏蔽”

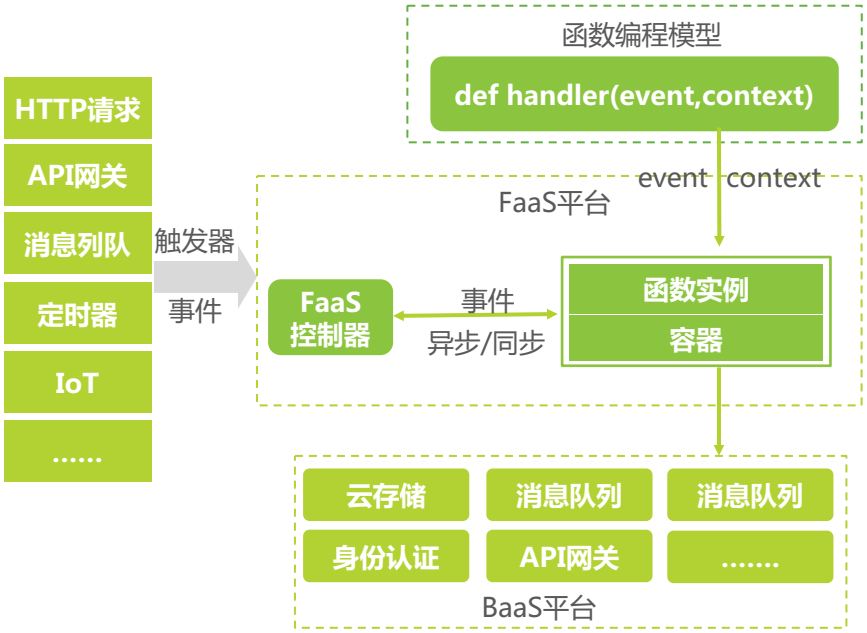
当前，微服务的生态和实践已经比较成熟，其设计方法、开发框架、CI/CD工具、基础设施管理工具等，都可以帮助企业顺利实施，然而其仍有许多不足：（1）粒度仍然比较大。（2）开发仍有较高门槛。（3）微服务基础设施管理、高可用和弹性仍然很难保证。（4）基础设施的成本依然较高。而Serverless中，开发者不再需要将时间和资源花费在服务器调配、维护、更新、扩展和容量规划上，这些任务都由平台处理，开发者只需要专注于编写应用程序的业务逻辑。如果再结合低零代码，则 “编写应用程序”的难度也大为降低，企业内的技术人员更加贴近业务。

微服务中，大量运维仍未被托管



来源：《华为serverless核心技术与实践》，艾瑞咨询研究院整理及绘制。

典型的serverless架构



来源：《华为serverless核心技术与实践》，艾瑞咨询研究院整理及绘制。

©2022.7 iResearch Inc. www.iresearch.com.cn

人工智能：需要大规模准确数据哺育

人工智能应用引发数据治理需求

企业在部署AI应用时，数据资源的优劣极大程度决定了AI应用的落地效果。因此，为推进AI应用的高质量落地，开展针对性的数据治理工作为首要且必要的环节。而对于企业本身已搭建的传统数据治理体系，目前多停留在对于结构性数据的治理优化，在数据质量、数据字段丰富度、数据分布和数据实时性等维度尚难满足AI应用对数据的高质量要求。为保证AI应用的高质效落地，企业仍需进行面向人工智能应用的二次数据治理工作。

AI应用对数据治理需求