技术驱动大数据价值释放

推动企业实现数据智能创新

本研究由阿里云发起

CONTENS
目　录

IDC观点 01

第一章
大数据产业发展现状和趋势特点 03

1.1 技术创新驱动大数据产业和服务发展 04

1.2 行业智能化应用促进大数据产业持续繁荣 05

第二章
大数据在行业的应用及挑战 07

2.1 机遇：各行业进一步加速数据价值挖掘 08

2.2 挑战：基础平台能力和数据智能水平亟待提升 10

2.3 破局：提升目标与应对策略，实现三个强化 12

第三章
构建企业大数据能力体系
推动全链路大数据实践 14

3.1 顶层设计，构建体系化和融合性方案 15

3.2 技术加成，着力打造高性能基础引擎 16

3.3 深化智能，将AI与数据智能深度结合 18

3.4 融合共生，全面拥抱云原生发展成果 19

第四章 最佳实践 22

第五章 IDC 建议 26

IDC 观点

大数据从互联网迈向互联网+

互联网企业是大数据技术发展的驱动者和实践者，其在推动全域数据治理的基础上，构建了一体化、全方位的数据服务，用数据驱动业务创新，进而实现高速发展。互联网大数据的深入实践，也为传统企业“弯道超车”提供了新解法，从生产到销售，从调度到运营，从服务到管理，大数据技术正在重新定义传统企业的未来。

从“夯实技术基础”走向“强化技术加成”，实现三个强化

无论是数据驱动发展的数字化原生企业，还是亟待降本增效、转型创新的传统行业，都已经或正在将深化大数据应用作为企业的核心战略。企业期望能够共享先进技术成果，加速数据价值的挖掘与释放，为企业创造新的增长动力。通过大数据领域的技术加成，**强化体系性，强化基础引擎性能，强化智能，**使企业大数据平台能够支撑更严苛、更复杂的需求：

- 满足大规模、复杂数据治理的需要
- 深入挖掘全量数据（特别是海量非结构化数据）的价值
- 不断提升高并发、低时延处理能力
- 进一步深化业务与人工智能的结合
- 释放大数据+云计算的巨大潜能

一体化和融合化成为大数据领域重要发展趋势

IDC认为，从提升企业中长期发展质量、降低综合投入成本的角度出发，大数据技术领域将呈现出两个显著趋势：一体化和融合化。

- **一体化：**为了构建快速响应业务的能力，大数据平台需要从较为单一的计算、存储、数仓基础设施，向具有广泛互通、互访、互操作能力的复合架构演进，以整合能力，降低成本，形成一体化的数据资源管控能力和高水平的数据全链路闭环流程。

IDC 观点

- **融合化：**大数据将加速与人工智能、云计算、区块链的结合，不断强化面向业务的实时、弹性和敏捷能力，同时对行业业务实践进行数字化积累与沉淀。例如，充分利用云上丰富的PaaS层中间件资源，保证系统的高弹性和高可用性；引入必要的SaaS化服务，构建低成本、敏捷化的处理能力；利用大数据加速算法模型的开发与迭代等。

着力构建体系化的大数据应用能力，驱动企业高质量发展。

企业应以战略和顶层设计为先导，用体系化的思维全面构建大数据能力架构，避免形成新的数据、业务和能力孤岛。IDC认为，企业可以从以下四个方面深入落实大数据实施策略：

- **顶层设计，构建体系化和融合性方案：**通过科学的规划、设计和实施方法，实现技术、产品、方案的融合；
- **技术加成，着力打造高性能基础引擎：**强化关键环节技术能力，以企业创新需求为导向，利用技术升级进一步强化大数据实时、敏捷、弹性能力；
- **深化智能，将AI与数据智能深度结合：**在企业分析、决策过程中充分引入AI能力，提供AI加成的数据，从系统层面将AI与大数据、BI等进行充分融合，实现良性迭代；
- **融合共生，全面拥抱云原生发展成果：**推广以云为基础的服务化模式，拥抱云原生领域的丰硕成果，全面提升企业IT服务的综合水平，实现降本增效。

01

大数据产业发展现状和趋势特点

1.1 技术创新驱动大数据产业和服务发展

随着企业数字化转型进程的持续深入，面向各行业的大数据服务已经开始向规模化、产业化方向发展。大数据领域的技术创新成果不断与行业业务实践相结合，催生了丰富的应用场景，形成了技术驱动产业发展的良好局面。

大数据市场的蓬勃发展源自于技术的快速迭代进步，这些进步又归因于各行业领域内持续增长的创新业务需求驱动。近年来，一些行业IT基础设施和应用系统呈现高速发展的态势，新增数据体量急剧膨胀，对数据处理性能、实时性和质量的要求也显著提高。

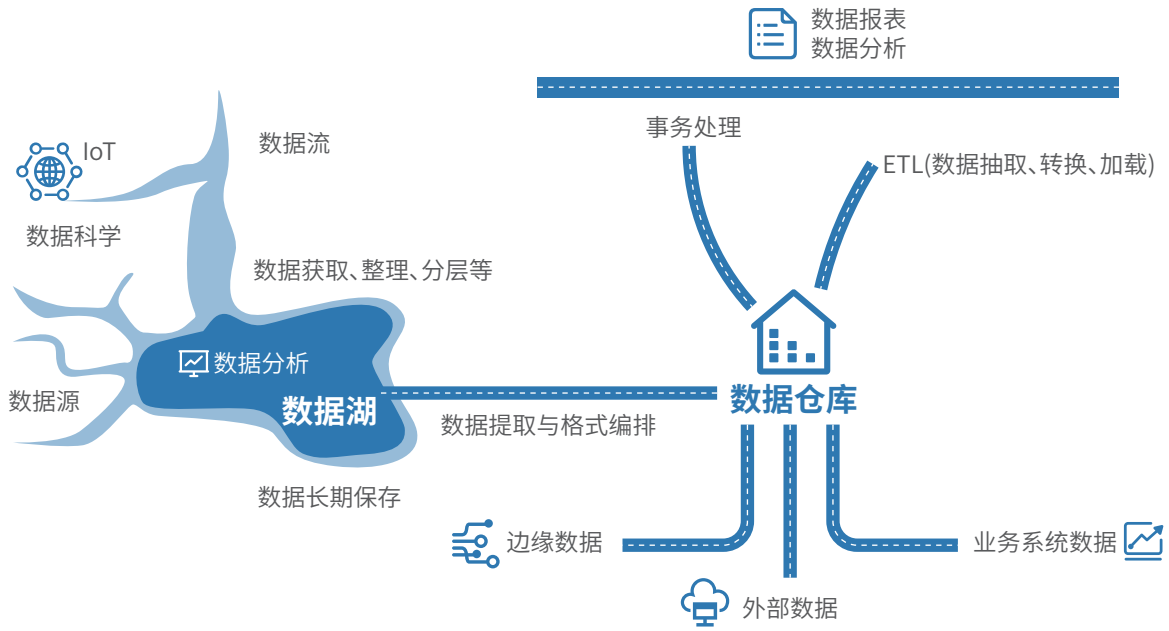
伴随过去十年的数字化浪潮，大数据技术架构已经经历了多次换代升级，从专注存储的数据库，到面向统计分析的数据仓库，再到面向大数据分析应用场景的大数据平台，各类大数据技术和服务商都在致力于通过**创新实践**，不断提升数据分析和应用能力。

数据基础创新：围绕大数据计算、存储、分布式处理等领域，追求更低成本、更强性能、更高可用性。例如海量数据分布式存储、快速全托管的在线数据仓库、数据湖、高性能离线计算与实时计算引擎等；

数据开发与治理创新：将方法论与工具相结合应用于智能化数据建模、交互式数据开发，全域数据集成与数据治理等，不断提升各环节的完备性和丰富度，形成完善的全链路数据治理与数据开发能力；

数据智能创新：涉及从大数据到人工智能的全链路AI工程化覆盖，涵盖数据标注、数据预处理、特征工程、算法开发、模型训练、模型优化、模型部署服务以及智能线上监控等全面能力，为企业高效利用人工智能创造一站式的开发环境，显著提升端到端落地的效率。同时，通过机器学习、深度学习技术，以场景化的能力推动企业业务创新，实现从数字化到数据化再到智能化的全面升级。

图1 数据治理关键技术和流程示意图

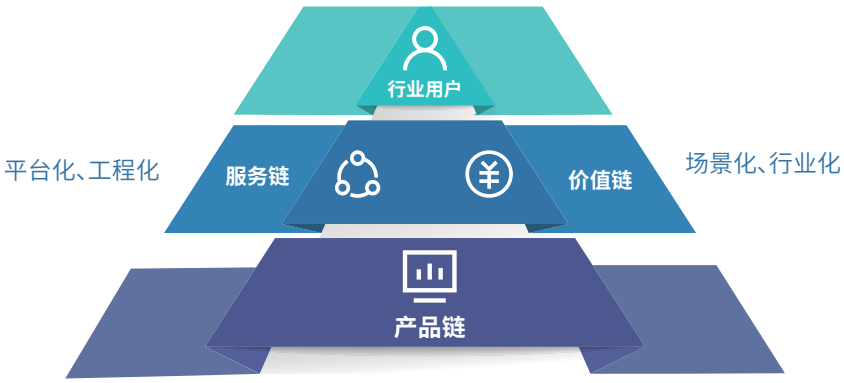


来源：IDC,2022

1.2 行业智能化应用促进大数据产业持续繁荣

快速增长的企业业务对大数据技术产生出强烈的渴求，推动了大数据产业的快速落地，发展形成了数据资源体系顶层设计、数据治理服务、数据平台建设、数据运维、数据开发生态等百花齐放的业态，大数据的产业化发展进程呈现出强劲的源头驱动力。得益于大数据产业开放、稳定、高效的服务特征，行业用户可以从健全的产品链、服务链、价值链中获取到高水平、高质量的能力支持。

图2 逐步形成稳定、高效的大数据产业链



来源：IDC,2022

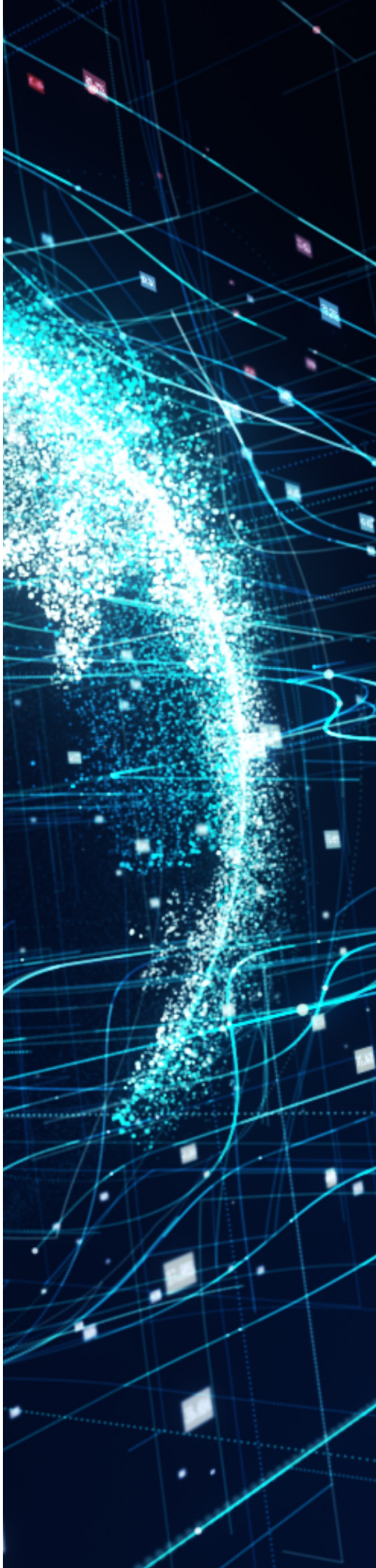
数字化原生企业带动传统行业用户，共同为大数据产业化发展提供充足的动力：

互联网等数字化原生企业是先进技术产生和落地的驱动者和应用者：在过去十年的大数据发展历程中，互联网企业作为数字化原生企业的代表，是大数据规模化应用的先行者，也是大数据先进技术的开发者和实践者。互联网企业具备先进的数字化业务平台，在产品开发、经营管理、市场营销、财务、运维等全领域，广泛采集和利用大数据赋能业务发展，取得了可观的成效与收益，大数据元素已经全面融入到这些数字化原生企业的血液中。一些大型互联网企业将大数据能力以及相关方法、经验进行长期沉淀，形成了丰富的理念、产品与解决方案。

传统行业为大数据发展带来更为广阔的应用空间：传统行业在加快数字化转型的步伐，很多企业已经步入数字化发展的纵深阶段，在企业运行状态监测、资源调度管理、生产与销售、经营与财务等领域，新一代的数字化系统正在源源不断地产生和积累大量数据，这些数据的潜在价值持续被认知和挖掘，推动大数据应用在核心业务中的占比不断提升。此外，云计算技术的快速发展，也进一步促使大数据的各项能力能够以服务化的模式，更广泛、更快速地被传统行业企业获取。综合各方因素来看，大数据对传统行业的深层赋能空间广阔，潜力巨大。

02

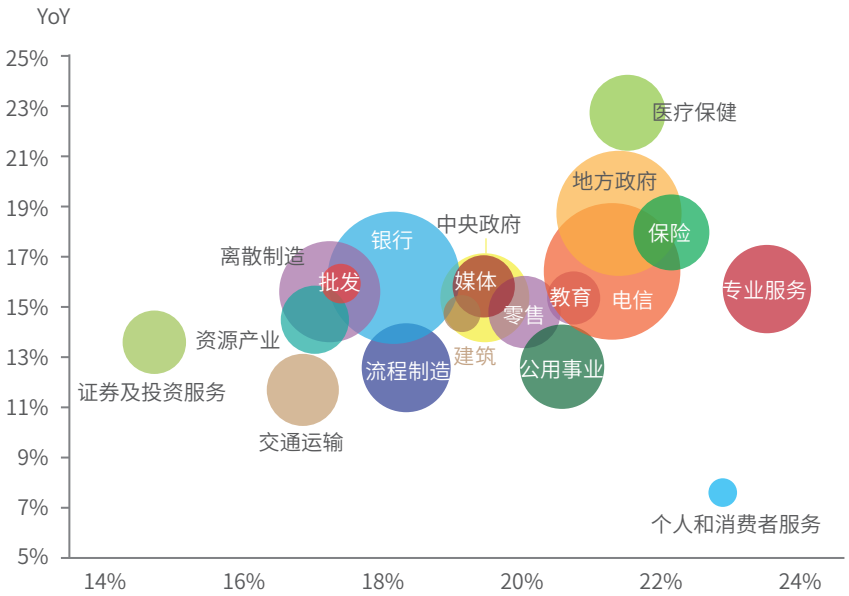
大数据在行业的应用及挑战



2.1 机遇：各行业进一步加速数据价值挖掘

IDC的市场研究和预测数据表明，无论是数据驱动发展的数字化原生企业，还是亟待降本增效、转型创新的传统行业，都已经或正在对大数据产生越来越多的诉求，各行业都期望能够加速数据价值的挖掘，为企业提供新的增长动力。

图3 中国大数据技术支出分行业市场规模预测，2020-2025年



来源：IDC,2022

我们按照大数据的能力发展阶段以及主要业务诉求，对不同行业进行了分类，以期望更清晰地认知不同行业对大数据规模化增长的预期。总体而言：

- 互联网、金融等领域技术能力较强，人才储备完善，正在持续深挖数据价值；
- 政府、交通、零售等领域建立了较好的基础，也具备了一批丰富的应用场景，但仍需持续积淀实践能力和经验，以进一步构建高阶数据智能能力；
- 能源、制造等领域需求旺盛，正从单点数据应用向体系化大数据应用迈进。

就行业发展特性而言，大数据的发展水平及发展重点，与行业的数字化成熟度以及业务发展要求密不可分。

行业	特征
互联网行业	主要业务大都会设立专门数据治理小组，推进全场景数据治理，在追求数据一致性、高数据质量的同时，更重视提升数据开发与计算的效率，同时做好成本方面的把控。
金融行业	普遍数字化转型成熟度和数据资源就绪度都相对较高，对数据安全的要求较为苛刻，大数据已经在诸如用户画像与精准营销、高频金融交易、小额信贷、智能检索、风控、交易历史数据查询等场景获得应用。
政府行业	数字政府大数据管理平台稳步增长，政务信息化逐步向融合化、智能化迈进，深化数据的跨部门、跨层级、跨区域的共享与应用，同时，诸如疫情防控等应用场景也要求数据的处理速度要更高，满足高并发场景下的稳定分析处理。
制造行业	越来越多的制造业企业在加速大数据技术的使用，利用物联网IoT进行数据采集，并探索大数据对于设备预测维护、流程优化、节能减排、供应链优化等维度的价值。但行业整体数据资源储备度不高，未形成专门的数据治理架构，目前多是面向一些分散的应用场景。
零售行业	零售行业数据准备度较高，可用数据规模大，来源广泛，诸如价格、品类、进销存、消费者行为等数据以及一些外部数据都可能被充分分析和挖掘，基于用户画像、商品画像、销售预测等模型，实现运营效率提升，支持业务拓展和创新，并优化生产管理决策，实践个性化营销。
能源行业	诸多能源企业近些年在大数据领域做了很多工作，有些机构的数据资源成熟度能达到DAMA 3-4 级的标准。部分企业在集团层面实现对财务、供应链、采购等方面的数据的统一管理，设定相应的专门数据治理部门，设立数据资源标准，但行业内企业数据就绪度不一，数据治理项目目前还比较缺乏统一规划。
交通行业	一大批与公交调度、信号控制、交通流采集、违规行为采集等相关的场景逐渐具备成熟的技术方案，推进城轨云和大数据平台相结合的落地应用，提升智能响应和自动化处理能力，并显著提高运营管理效率。在自动驾驶及辅助驾驶领域，相关专业大数据平台应运而生，为相关模型的开发迭代和测试提供强大的数据支撑，满足低时延、海量、高并发、地偏差等复杂数据处理需求。

2.2 挑战：基础平台能力和数据智能水平亟待提升

大数据在行业的价值和必要性逐渐被认可，应用前景越来越广泛，但IDC也注意到，企业内部的数据架构越来越复杂，业务诉求逐步在升级或细化，一些行业企业在具备了一定的大数据应用基础后，其大数据应用的规模和效用出现了较长时间的发展停滞。

IDC对大量行业企业的观察研究发现，企业在发展大数据应用的过程中，普遍面临以下四个维度的挑战：

缺少体系化发展路径，难以持续进步提升

越来越多的数据管理、分析、应用需求被纳入到大数据平台的建设中，企业大数据能力建设成为一项庞大的系统工程，涉及企业战略、业务流程、技术发展、组织与人才等多个维度的联系与影响。

在战略上：很多企业缺少数字化转型整体战略，缺乏统一的规划和顶层设计，体现在具体行动上，普遍以一两项紧迫需求为依据而随机上线某类大数据系统，很容易形成新的应用和数据孤岛。

在业务流程上：由于缺少科学的方法，没有建立业务与IT的协同机制，因此，原有的业务流程很大程度上会阻碍数据智能的采用。

在技术发展路径上：由于大数据分析系统规划滞后，分散建设，无法及时引入有助提升产品和服务实时性的新一代数字化技术，导致数据源质量差、数据标准不一致、数据项缺失、数据就绪度低等一系列问题。

在组织与人才队伍建设上：普遍不具备能与数据智能战略相匹配的组织体系，缺少既懂业务又有分析能力的人才，整体的大数据组织体系无法应对快速发展变化的内、外部需求变化。

关键环节能力不足，无法满足业务创新发展需要

许多快速成长的企业在业务创新过程中，产生了爆发式增长的海量数据，也对这些数据的并发处理和实时计算提出了很高的要求。例如：新兴的互联网/电商企业需要利用高效实时的数据处理，应对客户规模和产品品类快速增加所带来的精细化运营问题；一些传统的工业企业在持续推进IoT设备的规模化使用，并全力建设工业互联网基础设施，产生了前所未有的实时监测分析需求；此外，新一代智能交通/无人驾驶等领域的飞速发展，也正在对现有数据体系架构产生前所未有的冲击。这一系列新的变化，都对企业的大数据平台提出了更严苛、更复杂的需要：

满足大规模、复杂数据治理需要：企业数字化业务的快速增加，导致数据架构日趋复杂，而数据标准缺失、数据治理过程不规范、数据分析流水线过长等问题，需要企业重新审视在数据基础治理层面的能力缺陷；

挖掘海量非结构化数据价值的需要：随着各行业智能化场景的快速延申，大量与视频监控分析、图像关联分析、新媒体应用、文件处理等有关的业务领域不断出现，由此带来的大规模非结构化数据规模在急速攀升，需要创新性地解决海量非结构数据的存储、计算、管理、分析问题，形成综合性的解决方案，兼顾效率与成本；

应对复杂业务的需要：在直播电商、交通、公共服务、应急等领域，智能化发展趋势显著，大量新兴技术应用的落地和业务场景的产生，产生诸如交易数据、情感数据、位置数据、健康数据、互动数据、绩效数据、描述数据、行为数据不同内容的数据。而一些业务的峰谷效应明显，数据分析需求多样，需要可靠的弹性扩缩、智能增强以及一体化的服务输出能力，这对大数据平台的功能、性能都提出了十分苛刻的要求。

IT基础架构落后，疲于应对建设与运维难题

近十年间，IT技术（尤其是大数据技术）经历了日新月异的进步，企业IT架构也不断发生着深刻的变革。而很多企业的IT基础设施架构已经十分陈旧，无法有效进行升级，与当前的业务需求严重不匹配，各类系统的综合应变能力差，可靠性、可用性均无法得到保证。与此同时，由于设计思想落后，一些应用系统过于复杂，开发和运维人员始终被大量建设和运维问题困扰，无法将有限的精力聚焦于业务本身。纵观各行业的数字化转型进程，领先者与初入局者的差异正在被持续拉大，一些未能跟上数字化浪潮的企业正在面临无从下手的被动局面。

IDC观察到，许多企业从历史上延续下来的生产、经营以及内部管理系统，仍在发挥着重要作用，这些系统技术架构各异，服务接口缺失，难以被在线打通和整合，虽然这些系统内的一些关键数据存在潜在的巨大价值，但因采集成本和时效性的问题，无法对业务产生足够的正向反馈。此外，一些系统自身运行状态已经不堪重负，在业务量增加时，各类问题频出。

IDC认为，面向云化架构、承载新兴技术应用的未来数据基础设施是数字经济时代的必然产物，以云计算作为承载企业数字化优先战略的重要基石，必将加速企业数字化进程。

业务智能化水平长期停滞不前

一些企业的分析、决策和智能化加持的业务始终停留在较为粗浅的应用水平，难以使期待中的数据价值得到充分释放，这很大程度上归因于深度智能化应用能力的不足。在数据处理、分析和展示过程中，各类先进的机器学习和AI成果无法匹配到业务中，其根本原因是企业自身缺少智能化研究和发展能力，无法开发和积累相关领域的深度智能成果。

一方面，虽然基础数据量足够庞大，但有效数据不足：IDC数据圈研究表明，在今天创建的所有数据中，仅有大约27%的数据是被认为有用的，这其中44%被标记，21%被分析，15%在人工智能中发挥作用，这意味着在创建的总数据中，太多数据处于“沉睡”状态，约88%的数据还处于无组织环境中。

另一方面，大量的模型应用仍处于粗放的技术状态，很多模型有待进一步优化。总体上看，大模型是发展趋势，但企业仍缺乏符合自身业务发展和行业诉求的小模型。模型的训练依赖成熟、有效的数据，也需要不断根据新增数据来进行持续的优化和迭代。因此，数据治理能力与模型训练能力之间的开放互通，是进一步开发深度数据智能的关键环节。

IDC注意到：一些企业已经建立了专注于模型、算法开发的创新发展部门，这些组织在一定程度上提升了企业智能化应用的总体水平。但在日常的实践过程中，受制于自身理论水平、平台工具能力以及上下游部门的配合程度等因素影响，其整体工作成效和算法模型的产出物质量仍有待提升。

2.3 破局：提升目标与应对策略，实现三个强化

为了应对当前行业大数据发展过程中的多重挑战，面向企业未来巨大的创新发展空间，IDC建议从**目标与应对策略**两个方向深入分析问题，寻找破局挑战的方法。

就目标而言，IDC认为，大数据赋能企业数字化转型应涵盖四个维度：领导力转型、全方位体验转型、运营模式转型和智能化升级转型。要实现大数据支撑的全维度转型，需要从大数据的生命周期着手，建立健全“数据发现——价值开发——价值实现——知识与协作”的全生命阶段发展路径。针对目前企业在四个阶段面对的诸如数据就绪度、数据处理和价值挖掘、支持业务、实时敏捷、数据变现、智能化发展等挑战，IDC认为企业可以从企业大数据全生命周期的价值、应对方法等入手，思考相关的破局思路。

图4 企业大数据价值目标、挑战与破局思路

来源：IDC,2022

在具体应对措施上，应着重考虑采取必要的技术和综合管理手段，全面夯实基础能力，实现三个方面的强化。

- **强化体系性：**即充分认识到建设大数据体系化能力的重要性，通过科学的规划、设计和实施规程，实现各类技术、产品和方法的融合。同时，企业应建立起全局视角的数据管理和治理能力，实现对不同大数据引擎和异构数据存储的一体化管理，解决跨引擎数据开发、统一数据建模以及统一作业调度/管理等问题。

- **强化基础引擎性能：**通过深度介入底层技术，强化计算、存储、网络和机器学习等能力的综合性能。例如建立软硬一体联合优化的融合算力，大幅提高计算效率，甚至达到接近线性加速的能力。引入GPU虚拟化技术，实现最小虚拟化损耗。此外，还可以通过技术迭代升级，大幅优化IO性能、内部网络通信性能、大规模并行训练效率、推理效率等。底层技术的重构与优化，既可以大幅提升基础引擎性能，也能够在综合成本、敏捷性、灵活性以及体验等方面带来革命性改善。

- **强化智能：**在大数据平台深度植入AI能力，为企业带来AI加成的数据。即通过融合机器学习能力实现智能化分析、预测和模型匹配、优化。同时，在AI平台中拓展服务接口，让AI平台更便捷地处理分布式海量数据任务，降低AI的技术和人员门槛，提升大数据平台的全局影响力。

03

构建企业大数据能力体系，推动全链路大数据实践

3.1 顶层设计，构建体系化和融合性方案

数据资源将成为未来企业必不可少的发展要素，大数据应用在企业分析、洞察、决策过程中的影响力将逐渐由边缘上升为重要乃至核心，进而成为企业业务中不可分割的组成部分。企业有必要重新审视现有大数据系统的建设路径和产出成果，从企业发展战略的高度，客观认知现有大数据应用中的基础能力差距，并将规范化的大数据发展理论与企业实践相结合，用科学的方法建设面向未来的大数据应用体系。

形成体系化的数据治理与应用框架。传统的数据治理主要是面向 BI，不论是数据管理、元数据、主数据还是数仓建设，多使用单点工具组件，多面向单个应用系统。伴随行业大数据出现规模化、复杂化的显著趋势，各企业将致力于引入顶层设计措施，规划设计一套体系化的数据资源治理与应用框架，这一过程将涉及企业战略、组织、成熟度、标准、技术、产品、服务、创新体系等一系列规划设计要素。在这一趋势的引导下，企业应普遍从部署大数据单点产品发展为**构建大数据融合性方案**，综合考虑计算引擎、组合式存储、智能网络、全链路开发与治理、数据智能、数据多维展示等要素，形成符合企业未来发展要求的大数据体系架构。

规划设计全链路的闭环能力。通过良好的规划设计以及科学的建设方法，企业将有望建立大数据全链路的闭环能力，广泛覆盖企业大数据的采、建、管、用需求，并综合涵盖标准、汇聚、集成、治理、开发、运营、服务等过程。企业应统筹考虑资金投入与目标产出的关系，根据企业自身的实际状况，分步实施大数据建设计划，并在自建平台、混合架构平台，以及购买云服务等可选方式中做出综合选择，以更好地兼顾企业战略发展和现实经营需要。企业应能以新的体系化大数据平台为依托，实现对数据资源的一体化管控能力，全面管控好生产经营活动中产生的各类复杂、异构的数据资产，充分盘活数据资产的价值，最大程度地赋能到企业日常经营管理和业务创新环节，实现数字化驱动的、面向未来的全新业务模式。

参与大数据开发与应用生态共建。企业还需要提早布局数据开发与数据应用生态，这其中既涉及到与大数据服务商共建大数据服务产业生态，也包括企业通过大数据平台连通自身所在的行业生态的过程。因为企业一方面需要从服务生态中汲取先进的大数据技术经验，合理利用在先行企业中被证实行之有效的大数据技术方案体系，加速企业数字化能力的提升。另一方面，企业也可以通过先进的大数据平台聚集上下游厂商，在商业活动中不断提质增效，甚至能够基于大数据成果进行商业模式创新，为企业带来新的市场空间，进而带动全行业的健康发展。

3.2 技术加成，着力打造高性能基础引擎

为了构建快速响应业务的能力，企业会更加关注技术领域进步与提升，积极采用新技术成果，强化大数据的**实时、敏捷、弹性和智能能力**，特别是解决好**海量数据的高并发**和**实时处理**要求，为业务拓展和运营提供支持，这对企业的创新发展至关重要。

- 大数据计算、存储、处理、展示等基础能力的进步始终至关重要。随着工业互联网、自动驾驶、线上营销等新业务的大量兴起，企业需要紧密关注行业创新发展趋势，不断利用大数据强化分析、洞察能力，尤其是加强对海量非结构化数据价值的深度挖掘。事实上，大量的视频、图片、文件等非结构化数据已经被广泛应用在各行业的监测预警、辅助生产、事件处置等环节中，其数据规模和复杂度也已经超越了传统的结构化数据，成为未来大数据领域新的增长点。IDC预计，到2026年，非结构化数据预计将占IDC全球数据圈中90%以上的数据。

- 与此同时，业务端更多的新生业务场景也不再满足于传统的数据挖掘时效，大量的新业务要求对海量数据进行实时化、半实时化处理，并结合过往数据分析成果进行循环迭代。这些新的业务需求在很多行业甚至已成为刚需，例如金融的高频交易、风控系统，运营商的计费系统，政府的应急指挥、警务综合，交通的自动驾驶等等，海量的数据需要被快速的传输并处理，以满足越来越严苛的数据实时性需求。根据IDC预测，实时数据占比将从2019年的19.3%增加到2024年的24.4%。

今天的大数据技术呈现出分散化和多样化趋势，各领域细分技术快速变化，数据也可能来源于不同的数据库、数据仓库或者数据湖。**为了构建快速响应业务的能力，大数据平台需要从较为简单的计算/存储基础设施，向一体化、高性能引擎和云数智融合架构演进**，以满足跨系统、跨数据结构的复杂处理能力要求，建立从数据从生产到消费的全生命周期解决方案。在面对多维、复杂、异构的数据时，要综合考虑业务需求与技术支撑能力的平衡性，特别是处理好投入成本与预期成效的关系。

实现数据加工的敏捷化和实时化：随着在线数据服务引擎能力的不断突破，企业可以更好地满足自身在大数据分析过程中对敏捷性和实时性等维度提出的新要求。诸如分布式、向量化、列存储等现代技术架构使数据查询服务引擎的计算力获得了数量级层面的提升，受此积极因素的推动，传统需要逐层加工、逐层汇聚的数据处理方法开始发生变化。以保证数据质量为前提，在明细层或轻度汇总层即可对外提供交互式分析和探索式分析服务，既保证了数据的新鲜度、准确度、灵活度，也极大提升了数据研发的效率。数据开发方式从面向应用的烟囱式开发，转为面向数据资产和公共指标的主题开发模式，在数据服务层面，体现为减少副本的使用，通过描述性的SQL接口，以及高度优化的数仓引擎能力，强化数据服务层的统一和实时，降低资源占用，减少数据的不一致，也使IT团队与业务团队在一定层面上实现解耦。

实现性价比与效率的科学平衡：例如，将离线计算与在线计算的业务需求进行统筹考虑，精准地评估数据批量处理和流处理的需求比例关系，建立统一的底层支撑，实现更高效、更便捷的一体化管理模式。即利用一份物理数据同时对外提供离线与在线调用，减少数据的内部移动，使基础资源得到更加高效的利用，有效控制资源成本。此外，还可以通过技术升级，实现不同系统间元数据层的共享，使离线系统中新生成的表结构可以在在线系统里实时感知，减少元数据的不一致，显著提升引擎工作效率。

实现统一的数据加工逻辑：例如，从数据加工维度，将批量计算和流计算这两个不同的加工链路，融合为统一的加工逻辑，支撑多样化需求。这需要兼顾考虑离线计算的数据修正过程和在线计算的处理性能问题，甚至通过一套存储系统同时支持批量数据的海量高吞吐写入和流式数据的高频更新特性，对平台内部的开发和调度过程都有很高要求。企业在面对不同的计算要求时，只需要维护一套系统和脚本，既显著降低了使用成本，也有效避免了数据的割裂和不一致。

实现数据资产的持续积累：企业还应有序推进大数据高阶能力的建设，让数据资产成为业务创新发展的源生驱动力。依托数据资产的巨大价值，各业务部门能够不断推进业务细粒度监测分析、业务深入洞察、深度智能决策等多领域的进步，利用数据智能的强大能力提升精细化管理和精准决策水平。企业也将更快速地识别市场机遇，应对市场变化的挑战，与产业链、供应链实现高度协同。

案例

易仓将Flink计算后的实时消费数据通过Kafka写到实时数仓，并借助离线数据仓库与实时数仓的互通能力，完成复杂的分析计算，并通过SaaS方式服务全球租户，做到数据秒级响应。

3.3 深化智能，将AI与数据智能深度结合

AI与大数据的深度融合是未来大数据产业发展的重要趋势。海量数据的深层价值挖掘需要人工智能和机器学习手段的加持，而数据智能能力的进化与提升也离不开更加开放便捷的基础数据服务与互通。在大数据平台赋能业务实践的过程，企业会考虑将人工智能、机器学习领域的成果与大数据能力做深入结合，解决当前和未来业务领域的一些更复杂的问题。

目前，大数据与AI的结合呈现多种具体的形式。例如，一些领域的人工、半人工决策过程，将有可能通过引入AI能力而实现更大程度上的智能化决策，这有助于提升决策效率，减少决策过程中的人为干扰因素，同时也可以体现更细粒度的决策内容。在这一领域，一些AI增强式BI、智能决策平台等类型的产品已经得到初步应用。在模型开发和应用领域，一些行业领军企业依托长期的行业积累，生成了大量的成熟算法模型，这些模型在业务创新实践中不断被进行匹配、迭代和优化，并更多地通过人工方式加以管理。当引入AI能力作用于模型管理时，可以做到一定程度上的自动化算法识别、匹配和优化，甚至可以实现批量模型的生产与管理。

大数据+AI的模式将更多出现在企业纵深业务领域的创新活动中，作为大数据能力体系中更靠近应用层级的能力，其对相关的基础支撑技术有很高的要求。传统的AI并不适合处理分布式的海量数据，而大数据平台也缺少能够植入AI能力的服务，这不仅需要解决技术体系的融合问题，更需要通过平台将不同领域的人才连接起来，使他们为了一个统一的目标贡献力量。企业可优先在部分场景中引入AI与大数据一体化平台，重点考虑是否能真正提供适配本企业需求的AI能力，涵盖AI模型的开发、积累和迭代需求，甚至引入一些SaaS化的AI能力，快速解决企业的现实需要。在这个过程中，既要关注整体技术架构的灵活性，也要解决好数据的可获取性和一致性等底层问题，包括将计算、存储、网络等基础资源进行有效连接；打通数据、训练、资源管理、推理和服务；以及将AI能力进一步通过服务层的形式进行开放，使企业

更方便地获取标准化的算法模型与服务等。

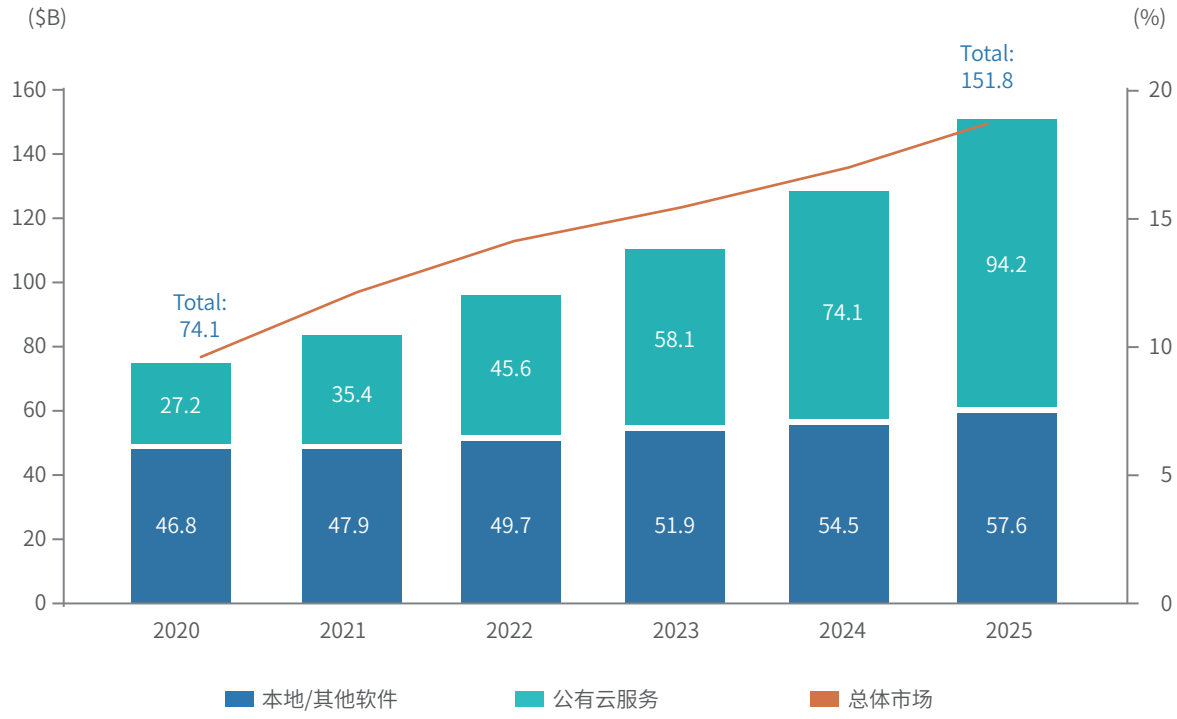
在强大的基础能力保障下，企业还要高度重视算法模型的创新工作。随着前端业务场景的日趋多样化，企业需要面向不同业务开发出更细粒度的算法模型，即具备数据分析模型的快速开发与修正能力，让不同部门的业务人员都能广泛享受到大数据带来的业务提升体验。由于大数据专业人员的普遍短缺以及对成本方面的考虑，企业将更多引入成熟的模型开发与管理工具，借助外部力量实现能力跃升。业务人员和数据运营人员能通过直观交互的平台化工具，与专业人员开展算法模型的共创，进而实现对模型的长期动态沉淀、管理和迭代优化，将企业运行过程中形成的数据能力和经验进行长期积累。

3.4 融合共生，全面拥抱云原生发展成果

各企业在发展大数据应用的进程中，还将不断提升以云为支撑的大数据基础能力，这与当前企业上云的趋势密不可分。在云上构建业务已经成为企业发展的一种战略选择，云的分布式架构和全栈云服务能力，能够帮助企业构建高可靠、高可用以及高性价比的丰富应用。同时，云所体现出的开放、共享、协同、智能的理念，有助于企业不断提升创新能力，更加快速地与行业领先水平对齐。

不断提升以云为支撑的企业大数据基础能力。云与大数据的深入融合，将为海量数据的处理和应用提供更高效、更高性价比的方案，大数据已经越来越多的与云技术、云原生密不可分，存算分离、统一管理，湖仓互访等技术发展趋势将不断推动大数据基础能力领域的深入挖潜，帮助用户打造更强的弹性和更好的一致性，有效应对突发业务变化，用更低的综合成本支撑企业长期发展。例如，将数据湖的开放式存储和多引擎能力，与传统数据仓库的管理和查询加速能力结合在一起，形成互通互访。这需要更好地利用云上开放存储、数据迁移工具以及安全访问控制的强大能力，让企业不同角色的数据分析任务变得更简单便捷。IDC预测，全球大数据平台公有云服务市场规模有望在2025年超过940亿美元，五年预测期内（2021-2025）实现约28.3%的复合增长率（CAGR），远超本地部署方式的复合增长率。2021年中国大数据平台公有云服务市场规模达33.7亿元人民币，相比2020年实现53.8%的快速增长。

图5 IDC全球大数据平台市场规模（单位：十亿美元），2020-2025年



来源：IDC,2022

积极推进系统迁移上云和云化重构。在云与大数据的融合过程中，企业可以结合系统迁移上云或云化重构的进程，对重要的大数据应用也同期开展云化设计工作，让业务应用全面拥抱云原生，充分享受云所提供的全栈能力和敏捷服务特性。在云上应用的设计中，可以充分利用PaaS层的丰富中间件资源，重构企业应用架构，让企业大数据获得更强大、更稳定的平台支撑。

结合业务需要更多地引入SaaS化服务，这是各类新兴企业特别是广大中小企业的理想选择。外部SaaS服务与企业自有业务系统建设过程相结合，能够使那些刚刚开启大数据之旅的传统行业企业，更快速、更便捷地获取到高阶数据智能能力。

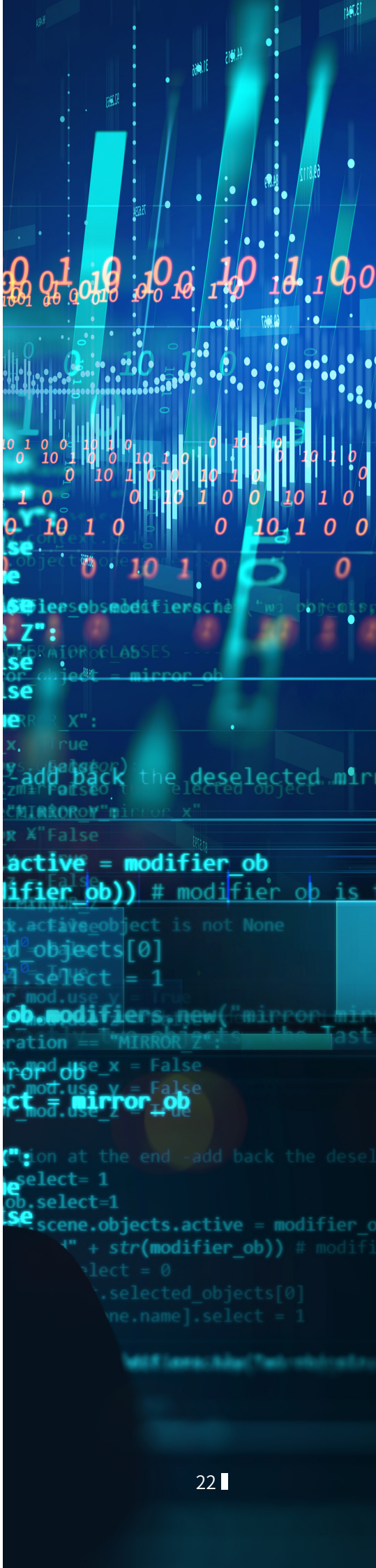
案例

快狗打车在业务快速发展的过程中，十分重视利用大数据促进精细化运营和智能化风控，目前已经建立完善的数据智能部门，涉及算法、分析、数仓、应用等一系列工作。快狗打车的大数据类服务已经全面上云，通过公有云上的数据同步工具、大数据计算服务以及实时数仓等丰富产品组件，构建了主流的云上体系架构，并根据业务特性，在批流一体、湖仓互通等方面进行了有效实践，利用大数据标签管理、画像分析等方法，对智能调度、智能定价、智能补贴、安全管理等重要业务环节提供服务，业务支持成效获得了质的飞跃，与历史同期相比，大幅提升了运营管理的精准度，也显著降低了综合成本。

未来三年，企业将继续强化云上数据服务的敏捷化能力，实现快速开发、快速交付、快速迭代，同时大幅降低底层维护的成本，将更多人力投入到上层的定制化、精细化应用场景中。

04

最佳实践



心动公司：

加速大数据与人工智能能力的融合，做更懂用户的游戏平台

心动公司在发展互联网游戏社区以及游戏研发、游戏发行、游戏运营等主营业务的过程中，全面利用大数据洞察各方客户需求、紧跟市场热点、优化运营流程、赋能内外部开发与管理工作。心动构建了规模化、体系化的大数据基础平台，积极推进大数据和AI之间的深度整合，并实现了数据湖和数据仓库的打通，切实发挥出了企业数据的巨大价值，也实现了企业长期的降本增效目标。

关于企业

心动公司创立于2003年，是一家全球游戏开发和发行商，拥有丰富的研发、发行和代理运营经验。截至 2022年6月，心动运营 38 款免费和付费游戏，在全世界拥有 5,000 万月活跃用户，主要分布在大中华地区、东南亚、北美和南美。2016年，心动推出手机游戏社区和应用商店 TapTap，截至2022年6月，TapTap 在全球有超过 5,000 万月活跃用户。

发展需求

作为一家互联网公司，心动公司始终以数据为驱动，挖掘数据价值，重点实现用户体验提升、改善运营效率。

- **洞悉玩家需求，精准快速服务：**TapTap基于对用户行为特征的洞察，及时发掘用户需求，以个性化推荐的方式提供千人千面的服务。
- **紧跟市场潮流，提供有价值的预测：**TapTap基于对市场数据的分析，迅速把握游戏市场的动态，掌握整体行业的发展趋势，提前判定可能会受市场广泛喜爱和关注的产品。

- **基于平台运营，优化匹配：**平台上的第三方游戏商希望将他们的游戏推荐到真正适合的用户手上，实现双方的精准匹配和高质量营销。

- **赋能游戏开发者，提升玩家体验：**基于细粒度的统计反馈结果，帮助开发者更好地了解用户习惯，不断提升游戏体验，同时能将游戏以及游戏相关的原创视频、文字等，更准确地送达理想的个人用户。

伴随业务向精细化、个性化发展，企业内部各方对数据分析服务的实时性要求越来越高，希望在用户访问之后很短的时间内，就能为用户推荐到更喜欢的内容。随着用户访问行为的深入，系统后台应具备快速学习能力，不断修正推荐策略，使用户体验获得不断进步。TapTap在早期选择自建大数据体系，随着规模的增长，自建平台的一些问题逐渐显现：

- **整体承载能力受限：**由于技术体系不够健全，平台存在一些影响长期发展的短板，其综合性能和稳定性欠佳，总体扩展空间有限。

- **技术保障能力不足：**自建平台需要维持足够规模的技术支持团队，由于总体投入有限，技术团队规模无法持续扩充，在跟进一些前沿的数仓技术或者机器学习技术时，无法支撑新场景的快速上线和快速迭代。

- **能力整合困难：**由于自建大数据平台和原有业务平台采用了不同的技术体系，产生了能力整合和生态对接问题，即无法在不同的基础设施平台上将业务和大数据做深度整合。

解决方案

心动旗下TapTap于2020年与阿里云合作，构建大数据+AI方案，支撑相关的搜索、广告、推荐、推送、投放等业务的进一步发展。

- **构建体系化大数据平台：**TapTap通过大数据基础计算引擎、大数据开发与治理服务平台以及高性能的实时数仓服务、实时计算服务、机器学习平台等，构建完整的大数据基础支撑和数据开发治理能力。

- **推进大数据和AI之间的深度整合：**基于阿里云大数据平台的一体化架构，TapTap可以实现数据计算与数据存储资源与机器学习平台的顺利对接，并获得模型训练、模型部署等维度的充分支持。

- **实现数据湖与数据仓的打通：**TapTap通过将离线数据和相关计算结果推到阿里云实时数仓，为线上业务做实时和准实时数据服务。同时，线上业务也会将实时计算引擎处理的数据写入到实时数仓里，将离线、在线数据整合在一起，统一对外提供服务。此外，企业还将早期建设的EMR数据湖接入到数据开发与治理服务平台中，与数据仓库打通，形成湖仓一体的初步成果。

应用成效

基于大数据体系化建设，阿里云大数据+AI服务的引入，以及湖仓数据的打通，TapTap获得更加充足的业务增长动力。

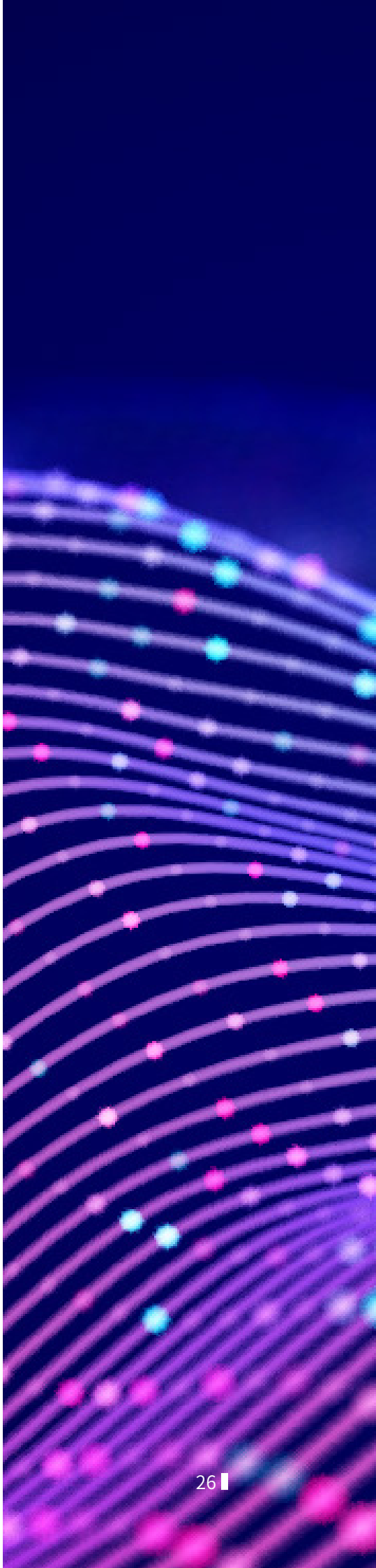
- **技术维度来说：**基于湖仓一体化的设计架构，提升数据应用效率，数据格式实现统一，为后续更广泛的应用奠定基础，也为更加实时有效的数据分析提供原料。
- **从业务角度来说：**对比先前，整个广告的匹配和推荐效率都有成倍的提升，在用户端，用户点击率以及游戏下载率也同样获得数倍增长，对应的广告收入获得大幅增加。

接下来，TapTap 会持续深化以下维度的尝试，以构建更加赋能业务的大数据能力：

- **提高模型的实时性：**考虑把模型的更新速度提到分钟级，同时持续提高模型的准确率；
- **提效降本：**综合考虑利用微服务架构重构部分搜推广服务，提高整个系统的弹性，降低综合成本；
- **AI与大数据的结合：**强化机器学习能力，将深度智能化应用在更多领域。
- **合规拓展数据服务：**尝试利用大数据实现联邦学习，希望在不侵犯用户隐私的前提下更多的与其他平台开展数据服务合作。

05

IDC建议



理论引导，整体规划，分步实施

企业需要以科学的理论做指导，在实践中建立与大数据相关的战略、文化理念、管理方法、平台和交付/运维体系。传统企业更应体现后发优势，重视整体规划工作，根据自身的实际发展状态，依托统一的战略，分步构建大数据应用体系，避免在大数据赋能业务的过程中，形成新的系统和数据孤岛。企业的大数据体系化能力，是未来长远发展的重要基础要素，需要得到企业内部上下一致的重视与认同。

迈向行业纵深，形成长期积累

在大数据实践过程中，企业既要聚焦业务需求，挖掘多样化的大数据应用场景，更要注重数据科学与行业知识的融合，建立持续迭代的行业大数据发展能力。企业应通过高水平的平台建设，降低开发人员的技术门槛，减少资源冗余，进而最大程度地降低综合投入成本。企业还应努力依托平台所能提供的丰富功能，形成规范有序的数据资产，积累和迭代多样化的分析模型，以此对业务能力和经验进行长期沉淀。

技术引领业务创新，全力提升应用体验

企业大数据应用是典型的创新型和技术密集型发展领域。基础技术的每一次升级迭代，都会带来性能的成倍提升以及综合成本的大幅下降。企业需要对大数据领域的发展动态保持足够的敏感性，洞悉新技术发展走向，在业务创新的过程中取得适度的技术领先，进而转化为业务效率和经营水平等方面的竞争优势，不断增强发展韧性与活力。通过基础技术领域的革新，企业有望在大数据的各子领域里全面深化基础能力，显著提升内、外部各使用方的应用体验。

适时、合理利用先进商业化成果

企业应坚持有所为有所不为，将有限的精力聚焦在业务创新活动上。在大数据能力构建过程中，注重利用先进领域的商业化成果，建设专业的数据基础工程。在大数据技术方案的设计中，应兼顾考虑技术路线、产品选型、模块化搭建和多维服务能力构建等关键要素，不断优化企业IT人员的使用环境。在服务商选择上，应注重服务商的整体咨询能力，关注其是否能够通过有效的咨询服务帮助企业建设体系化的大数据工程，并重点强化关键环节技术，保证系统长期、稳定、高效运行。