中国大数据分析行业研究报告

中国大数据网
2022 年 4 月

声 明

本研究报告针对的是中国的大数据分析市场，研究重点主要聚焦在新兴型厂商。中国大数据网结合自身数据库信息，并采取访谈调研、专家研讨等多种方式，对大数据分析行业涉及到的多个细分市场进行定量和定性的分析，给出观点和结论，以供政府机构、科研机构、企业和产业投资机构参考。

本报告中使用了新兴型厂商 2020 年度的主要营业收入对其市场份额进行排名，中国大数据网将在有关单位进行 2021 年度数据更新后对本报告涉及的部分数据进行调整。由于数据来源、模型设计、调研访谈和专家研讨等环节的局限性和差异性，报告难免存在不足之处，欢迎各界讨论指正。

本报告发布渠道：

“中国大数据网” www.zgdsj.org.cn

“中国科技新闻网”www.zghy.org.cn

“中国大数据网”微信公众号、“中国科技”微信公众号

目录

图表目录

1 研究背景

在全球信息化快速发展的大背景下，大数据已成为国家重要的基础性战略资源，正引领新一轮科技创新，推动经济转型发展。紧密围绕数据资源开展的基础设施建设、数据集聚整合、数据分析处理、数据开放共享和数据安全，铸就了大数据产业发展的核心要素。这些要素所构筑的“内层齿轮”的转动直接带动了“外层齿轮”——大数据融合应用的蓬勃发展，衍生出政府大数据、互联网大数据、健康医疗大数据、金融大数据、电信大数据和工业大数据等热点场景，持续驱动经济增长和转型升级。

“十三五”时期，我国大数据产业快速起步。据测算，产业规模年均复合增长率超过 30%，2020 年超过 1 万亿元，发展取得显著成效，逐渐成为支撑我国经济社会发展的优势产业。

政策体系逐步完善。党中央、国务院围绕数字经济、数据要素市场、国家一体化大数据中心布局等做出一系列战略部署，建立促进大数据发展部际联席会议制度。有关部委出台了 20 余份大数据政策文件，各地方出台了 300 余项相关政策，23 个省区市、14 个计划单列市和副省级城市设立了大数据管理机构，央地协同、区域联动的大数据发展推进体系逐步形成。

产业基础日益巩固。数据资源极大丰富，总量位居全球前列。产业创新日渐活跃，成为全球第二大相关专利受理国，专利受理总数全球占比近 20%。基础设施不断夯实，建成全球规模最大的光纤网络和

4G 网络，5G 终端连接数超过 2 亿，位居世界第一。标准体系逐步完善，33 项国家标准立项，24 项发布。

产业链初步形成。围绕“数据资源、基础硬件、通用软件、行业应用、安全保障”的大数据产品和服务体系初步形成，全国遴选出 338 个大数据优秀产品和解决方案，以及 400 个大数据典型试点示范。行业融合逐步深入，大数据应用从互联网、金融、电信等数据资源基础较好的领域逐步向智能制造、数字社会、数字政府等领域拓展，并在疫情防控和复工复产中发挥了关键支撑作用。

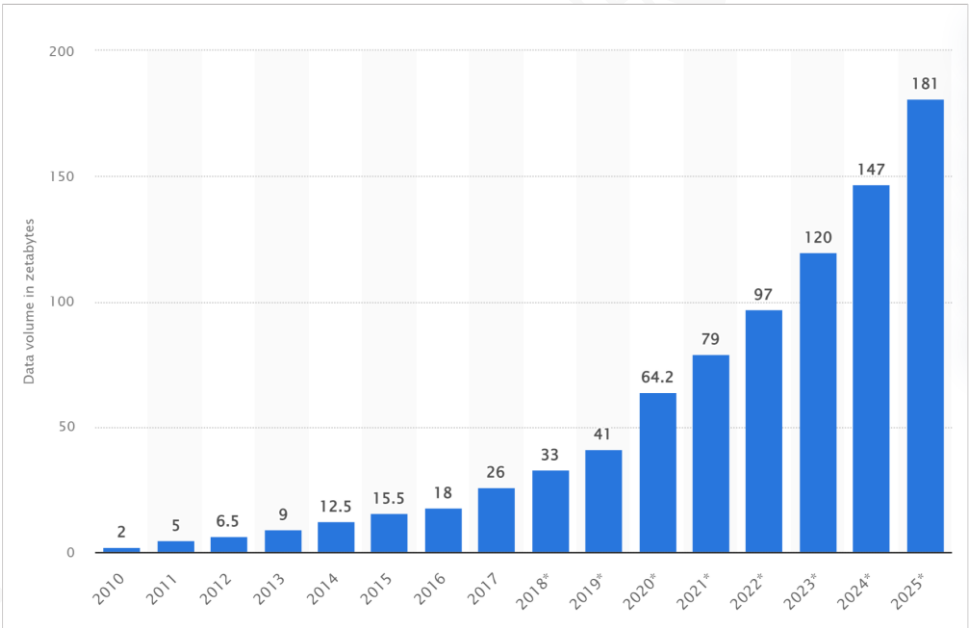
生态体系持续优化。区域集聚成效显著，建设了 8 个国家大数据综合试验区和 11 个大数据领域国家新型工业化产业示范基地。一批大数据龙头企业快速崛起，初步形成了大企业引领、中小企业协同、创新企业不断涌现的发展格局。产业支撑能力不断提升，咨询服务、评估测试等服务保障体系基本建立。数字营商环境持续优化，电子政务在线服务指数跃升至全球第 9 位，进入世界领先梯队。

中国大数据产业发展受宏观政策环境、技术进步与升级、数字应用普及渗透等众多利好因素的影响，市场需求和相关技术进步成为大数据产业持续高速增长的最主要动力。中国大数据产业呈现集成创新和泛在赋能的新趋势。新一轮科技革命蓬勃发展，大数据与 5G、云计算、人工智能、区块链等新技术加速融合，重塑技术架构、产品形态和服务模式，推动经济社会的全面创新。各行业各领域数字化进程不断加快，基于大数据的管理和决策模式日益成熟，为产业提质降本增效、政府治理体系和治理能力现代化广泛赋能。随着“互联网+”的

不断深入推进以及数字技术的不断成熟，大数据的应用和服务持续深化，中国大数据产业将继续保持高速增长，创新力强、附加值高、自主可控的现代化大数据产业体系将基本形成，持续促进传统产业转型升级，激发经济增长活力，助力新型智慧城市和数字经济建设。

2 大数据产品概念和分类

大数据是数据的集合，以容量大、类型多、速度快、精度准、价值高为主要特征，是推动经济转型发展的新动力。



图表 1、全球生成、获取、复制、消费的数据量（单位 ZB），Statista 2022

大数据应用的蓬勃发展是社会进步的必然结果，互联网普及之后，数据的生成、获取、复制、消费呈现出指数级发展的趋势，这些数据来自气象卫星、交通摄像头、车联网、电力/能源/工业/环保行业的数控设备和传感器、社交媒体动态、音频视频消费习惯、移动应用的

用户使用行为、购物平台的浏览和购买记录、服务器的事务记录及安全日志等等，企业和政府利用这些数据制定决策，完善流程和政策，并打造以用户为中心的产品、服务和体验。通过挖掘和分析这些数据，企业能够提高自身的竞争力和抗风险能力，把握新机遇，革新业务模式；政府能够洞察趋势、制定出更科学的决策和政策。在现代社会环境下，不进行大数据分析，就会“耳聋眼瞎”。

2.1 大数据发展的驱动力

大数据在业务需求和技术创新的结合中蓬勃发展。许多以大数据为核心战略的公司取得了巨大的成功，阿里巴巴、腾讯、亚马逊、苹果都是大家耳熟能详的例子。大数据的迅速发展有六个主要的驱动因素：社会数字化、物联网、技术成本快速下降、云计算的快速发展、数据科学的普及、人工智能的崛起。

1. 社会数字化

消费者驱动和面向消费者的大数据是最重要的大数据应用，“永远在线”的消费者生产了大量的数据。

据 2021 年 2 月中国互联网络信息中心(CNNIC)发布的《中国互联网络发展状况统计报告》，中国有 9.9 亿网民，平均每天的上网时间为 3.7 小时。网民与网民之间、网民与企业之间、网民与政府之间的每一次交互，每次点击、滑动或处理消息，都会在各大平台的数据库中创建新数据，巨大的人口基数创建出了巨量的数据。

新冠肺炎疫情加速推动了从个体、企业到政府全方位的社会数字

化转型浪潮。疫情的隔离使个体更加倾向于使用互联网连接，用户上网意愿、上网习惯加速形成。网民个体利用流媒体平台和社交平台获取信息，借助网络购物、网上外卖解决日常生活所需，通过在线政务应用和健康码办事出行，不断共享互联网带来的数字红利。在企业方面，疫情的出现为企业数字化转型按下了“加速键”，在线办公、在线交易等线上化运营方式为企业在特殊时期保持正常运转提供了支撑。在政府方面，政府的数字化应急能力和在线政务服务能力在疫情下不断“淬炼”，在线服务指数由全球第 34 位跃升至第 9 位，迈入全球领先行列。

2、物联网

物联网（IoT）是指通过传感设备、按照标准通讯协议，把物品与互联网连接，实现智能化服务的网络。物联网大致诞生于 2008 年到 2009 年，互联网上连接的物的数量超过了人的数量时，物联网诞生了。工业、商业和公共设施领域很早就开始了物联网应用，智能汽车市场也开始爆发，一辆普通家用轿车上大约有近百个传感器，而且传感器数量还在不断增长之中，更有大量的智能设备开始走入家庭，家庭里的 IP 地址数量急剧增长。据 Business Insider, IoT Analytics, Gartner, Intel, Statista 联合预测，2025 年全球将有 309 亿设备接入物联网。这些物联网设备能够以更高的速率不知疲倦地生产数据，对这些设备的运营、监控以及安全保障，离不开大数据技术的支撑，反过来也推动了大数据技术的进步。

3. 技术成本快速下降

大数据相关的技术变得越来越便宜，数据存储和处理的成本不断下降，使小型企业甚至个人都能够参与大数据应用。摩尔定律在大数据领域适用，计算能力的性价比每两年翻番，存储密度以及容量也每两年翻一番。相较于世纪之初的 2000 年，在造价相同的情况下，现在我们可以获得超过 1000 倍的计算性能和超过 1000 倍的存储容量。

除了计算和存储成本的下降之外，影响大数据系统成本的另一个关键因素是开源大数据软件。与价格高昂的数据仓库时代相比，这些开源软件以及基于开源软件快速成长、扩散的技术服务能力，极大地降低了大数据项目的成本。

4. 云计算的快速发展

云计算以及云计算环境下大数据技术的成熟，使构建一套大数据系统从高投入、高风险的项目（需要大量专家长时间进行系统搭建），变为低门槛、快速启动的项目（基础软件可以在若干分钟之内搭建完成），而且能够随着业务的增长进行无缝的技术增长，只需为实际使用的计算和存储资源付费。云计算大幅降低了大数据系统的技术门槛、时间成本和使用成本。

5. 数据科学的普及

新世纪以来，数据科学和数据科学家这两个词变得非常流行。《哈佛商业评论》称数据科学家为“二十一世纪最性感的工作”。近年的职场上，对数据科学家和类似职位的需求急剧增加，许多人积极投身到数据科学领域。对数据科学的教育更加专业化，统计和数据分析专业，正在变为学生和工作人群中的热门专业。数据科学的普及为大数

据的发展贮备好了智力资源。

6. 人工智能的崛起

大数据发展的早期阶段，具有数据科学知识是进入大数据行业的基本门槛。进入 2020 年代，随着人工智能带来的革命性变化，数据分析的技术门槛已经大大降低，大量未经数据科学训练的业务人员也可以方便地使用大数据系统了，实现了数据分析“平民化”的效果，大大促进了大数据技术的应用深度和广度。

2.2 大数据产品分类

图表 2、大数据产品分类

大数据市场经过 20 多年的长足发展，形成了丰富的市场生态，从产品、服务供应端的视角大致可分为如下领域：大数据基础设施、

大数据分析、大数据应用、大数据开源项目、数据源和 API、数据资源。下面我们对各个领域择要讨论，并对数据分析相关部分着重展开。

2.2.1 大数据基础设施

大数据技术的高速发展期开始于本世纪初，其前身是数据库技术。随着数据规模持续的高速增长，主流的技术覆盖范围从“数据”变成了“大数据”，其基础技术的演化大致有如下脉络：

1、 数据库

2、 数据仓库

3、 数据湖

4、 湖仓一体

这些技术彼此之间并没有淘汰或取代的关系，他们各自有自己的定位和擅长的业务场景，共同构成了大数据时代的技术基础设施。

数据仓库是个诞生于数据库时代的概念，早期服务于超大型企业的决策支持，并且也在不断地与时俱进，云上数据仓库服务获得了更多的大中小型客户，是对结构化数据进行分析的大数据技术。

数据湖源自于大数据时代开源技术体系的开放设计，经过云计算服务商的积极推广，在新兴公司中大量被采用。通常是由一系列云产品或开源组件共同构成大数据解决方案，可以处理一系列格式不同的结构化、半结构化、非结构化数据。

数据仓库和数据湖是大数据架构的两种设计取向，两者在设计的根本分歧点是存储系统访问、权限管理、建模要求等方面的不同。

数据湖和数据仓库作为大数据系统的两条不同演进路线，有各自特有的优势和局限性。数据湖对初创用户友好但成长性不佳，而数据仓库则刚好反之，对初创用户不友好但成长性较好。

历史较长的用户一般都成长于数据库时代，数据仓库（如果有建设需求的话)是当时唯一的选择。进入互联网时代，云上的半结构化、非结构化数据越来越多，也需要进行处理的时候，传统的数据仓库就遇到挑战。

相当一部分新型企业（尤其是新兴的创业公司）从零开始架构的大数据技术栈，正是伴随开源大数据软件的流行，天然地选择了数据湖架构。随着业务的不断发展，数据湖架构的问题开始显现，它太过灵活而缺少对数据监管、控制和必要的治理手段，导致运维成本不断增加、数据治理效率降低，企业落入了“数据沼泽”的境地，即数据湖中汇聚了太多的数据，反而很难高效率的提炼真正有价值的那部分。

湖仓一体的架构应运而生，兼顾数据湖的灵活性和数据仓库的成长性/事务性。湖仓一体的实现路径有两种。第一种，在数据仓库上支持数据湖，一般方案是在数仓中建外部表；第二种，在数据湖中支持数仓能力，一般方案是做一些开发，比如多版本并发控制、自适应schema、提供文件级事务等等。两种实现路径都需要解决一些共性问题，如数据打通问题、元数据一致性问题、湖和仓上不同引擎之间数据交叉引用的问题、湖仓开发工具缺乏问题等等。

湖仓一体的架构是新一代大数据分析的基础设施。

2.2.2 大数据分析

大数据分析领域有商业智能平台、可视化、数据分析师平台、增强分析、数据目录与发现、指标平台、流批一体、日志分析、查询引擎、搜索等细分领域。

2.2.2.1 商业智能

商业智能（BI，Business Intelligence）是大数据分析最典型应用领域，指以大数据基础设施系统为基础，运用各种数据分析手段进行数据分析以实现商业价值，部分商业智能的输出结果会以可视化的方式展现。

商业智能不是严格意义上的一种技术，它是数据库、数据仓库、数据湖、湖仓一体、ETL、OLAP、数据挖掘、机器学习和人工智能、资料展现等技术的综合运用，把它视为一套配合业务的流程和解决方案更为合适。

商业智能的关键是从许多来自不同的数据源中提取出有用的资料并进行清理，以保证资料的正确性，然后经过抽取（Extraction）、转换（Transformation）和装载（Load），即 ETL 过程，合并到数据仓库里，从而得到企业资料的一个全局视图，在此基础上利用合适的查询和分析工具、数据捕捞工具、OLAP 工具、机器学习和人工智能技术等对其进行分析和处理（这时信息变为辅助决策的知识），最后将知识呈现给管理者，为管理者的决策过程提供支持。人工智能在商业智能里开始扮演越来越重要的作用。

2.2.2.2 数据可视化

数据可视化把抽象的数据以人类容易理解的形式进行展现，常见的展现形式包括：图形图像处理、计算机视觉以及用户界面，通过表达、建模以及对立体、表面、属性和动画的显示。数据可视化可以大幅度提高人们对数据涵义的沟通效率。

2.2.2.3 数据分析师平台

数据分析师通常来自业务领域（相当一部分是商业智能系统的用户），通过洞察数据发现背后的业务趋势，数据分析师使用的最经典的工具可能是 Excel 电子表格，以图形化的方式操纵各种工具获得结果。

数据分析师平台正是这种易于使用的图形界面平台，不要求用户具备编程能力，大大降低了数据分析师的人员技术门槛，使人们更多的精力投入到业务领域。

数据分析师平台通常具备对各种格式的原始数据进行数据转换的能力，支持工作流，支持简单代码或无代码处理方式，可以直接输出结果进行展现，或者把处理结果输送到更复杂的工具中进行进一步处理和展现。

2.2.2.4 增强分析

增强分析是指使用机器学习和人工智能等提升能力的技术来协助进行数据准备、洞察生成和洞察解释，从而增强人们在分析和 BI 平台中探索和分析数据的能力。

增强分析可以将内部数据与外部数据相结合，并自动执行重要且

耗时的任务，例如数据准备、可视化、预测和报告。使用机器学习的增强分析平台，可以使数据分析更智能、更准确。技术是自动化和增强的，可以更快、更智能地获得对所有数据可视化、企业报告、场景建模和移动分析的洞察力。

增强分析中应用了人工智能技术，通常以机器学习(ML)和自然语言处理(NLP)的形式嵌入到分析中。它与传统的分析或商业智能(BI)工具有很大不同，因为机器学习技术始终在幕后工作，以不断学习和增强结果。增强分析可以更快地访问从大量结构化和非结构化数据中获得的洞察，并提供基于机器学习的建议。这种智能有助于发现数据中隐藏的模式和偏差，消除人为偏见，并启用预测能力来告知组织下一步该做什么，引导用户发现他们原本无法看到或发现的洞察。

增强分析的价值具体体现在如下三个方面：

- AI 使得大量的业务人员快速获得数据分析能力，不需要数据科学的专业知识，也不需要技术人员的支持，而且数据的使用也在统一的数据架构和安全架构之下，在大大降低了使用者的技术要求之后，业务人员更容易获得数据之下的业务洞见。
- AI 可以使用自然语言与人交互，并在交互中进一步学习，可以对数据洞察进行个性化处理。由自然语言处理(NLP)和自然语言生成 (NLG)组成的自然语言界面(NLI)，使用户可以用简单的语言提出问题并以简单的语言得到答案。用户能够使用直观的探索工具更深入地了解他们的数据。在用户问题的指导下，系统会推荐可视化图表、仪表板和其他易于理解的指标，展现出令

人信服的数据。

- AI 可以自动地进行数据清理和准备，自动完成繁琐的数据准备工作，让 IT 工程师和业务分析人员能够更高效地从事他们的本职工作。

人工智能（AI）是指计算机系统模仿人类的认知活动，能够“思考”和解决问题，并且不断学习进步。机器学习是人工智能的子集，利用数学模型和大量的数据来生成新的认知，不需要人类告诉它规则，它可以从数据中找出规则。机器学习是计算机的“智能”能够不断进步的根本原因。

机器学习与传统编程有极大的不同，在传统编程中，我们按照既定规则来编写代码，接收数据输入，然后产生正确输出。但对于许多认知智能领域的任务来说，制定规则是十分困难的。例如，区分是猫还是狗对人类而言是很轻松的任务，但描述其区分规则却相当困难，更不用说把它变成程序代码了。

图表 3、传统编程与机器学习模型对比

而机器学习另辟蹊径，它从一些输入数据和正确的输出开始（“图 1、2、3 是猫，图 4、5、6 是狗”），以此为基础的机器学习算法会生成规则，包括人类不知道的规则，这些规则汇聚在一起称为

机器学习模型，经过足够大数据量的训练之后，机器学习模型就能够有效地反映现实世界中的规则了（可以有效地区分猫和狗了）。

换句话说，机器学习通过一组自定义的学习规则分析复杂的数据集来增强模型。机器学习模型从大数据和重复的人类交互中学习，直到它可以输出足够好的结果。

随着数据的极大丰富、算法的不断进步和机器算力的大幅提升，人工智能在部分智能领域近年已经达到或超过了人类的能力，到达了“可以用了”的水平。

机器学习可以从数据中构建出规则，这正是历史悠久的数据分析工作梦寐以求的目标。人工智能与机器学习在大数据分析领域中开始扮演越来越重要的角色，也代表着未来，它在商业智能和增强分析中已经成为不可或缺的部分。

2.2.2.5 数据目录与发现

数据目录是关于数据资产的一个有序清单，它使用元数据来帮助组织管理数据，帮助数据专业人员收集、组织、访问和充实元数据，从而为数据发现和治理提供支持。数据目录之于数据，正如图书目录之于图书。它可以提供一个整体视图，提供所有数据的深度可见性，而不仅仅是一次只查看某一项数据。

与过去相比，想从如今前所未有的数据海洋中找到正确的数据更加困难。同时，关于数据的监管条例和法规也比过去更多、更严格。在这一背景下，除了数据访问之外，数据治理也成为了一个严峻的挑

战。不仅要了解当前所拥有数据的类型、哪些人在移动数据、数据的用途以及如何保护数据，还必须避免过多的数据层和封装，避免数据因太难使用而毫无用处。

数据目录可以使用元数据来实现比传统数据管理更丰富、更强大的功能。

2.2.2.6 流批一体

流批一体是指将流式计算与批量计算两种不同架构的数据处理模式融合到一起。

流式计算与批量计算模式的选择，是由用户使用场景决定的。流式计算适合于有实时或准实时需求的场景，将数据流连续地送入分析工具并快速地得到分析结果，如欺诈实时检测、社交媒体情感分析、安全日志监控、客户行为分析、实时推荐等；而批量计算则适合于非实时的场景，将一段时间内产生的大块数据一起送入分析工具，经过较长运行时间得到结果，如工资单计算、计费、客户订单、清算对账、指标分析、离线报表等。下表对比了两种计算模式的不同：

特性	批量计算	流式计算
数据时间范围	有界数据集，数据在某个时间范围内起始和结束	无界数据集，一直有持续不断新产生的数据
任务执行	分批执行、有终止	持续执行、无终止
处理延迟	小时级、天级	秒级、分钟级

数据场景	数据量超大数据、无法以流的形式交付	数据以流的形式交付
业务场景	工资单计算、计费、客户订单、清算对账、指标分析、离线报表	欺诈实时检测、社交媒体情感分析、安全日志监控、客户行为分析、实时推荐
关注点	可扩展性、大吞吐量、容错	可扩展性、低延迟、容错、消息一致性、消息持久性

图表 4、批量计算与流式计算对比

对于用户而言，只要数据量达到一定规模，对流式计算和批量计算就会产生业务需求，两种模式需要同时存在，随之而来的是一系列问题：

- 重复的资源，存储和计算都要双份，系统的成本高。
- 两套系统，组件不同，需要技能不同的人员维护，人员的成本高。
- 两套开发体系无法统一，表结构不同，开发环境不同。
- 缺乏数据一致性，对于相同的指标，两种模式算出来的结果不一样，虽然最前端输入都来自同一份源数据。

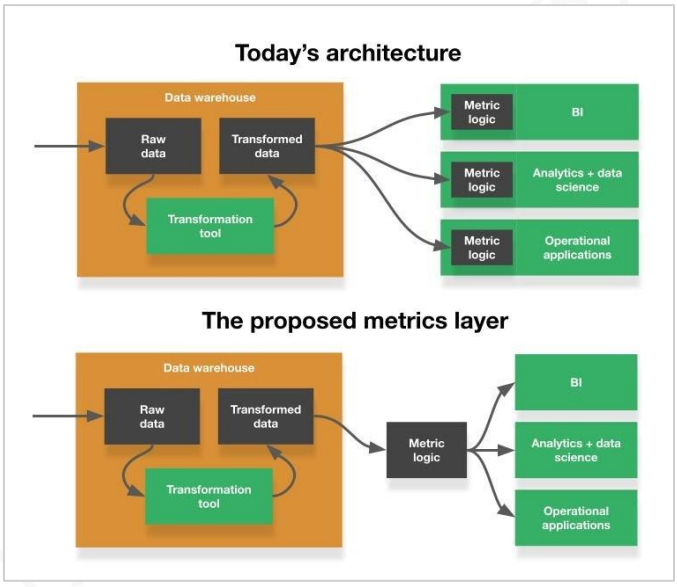
为了解决以上问题，流批一体成为新热点，目标是建立起一套统一的架构，可以同时支持流式计算和批量计算，对混合的有界数据和无界数据能够统一进行支持，提供更一致的、更广泛的编程环境，以较少资源浪费，降低维护成本，获得更好的数据一致性。目前，流批

一体的产品和服务正在快速发展和完善之中。

2.2.2.7 指标平台

这里的指标（Metrics）是指业务上或技术上关注的量化信息，例如销售部门关注的销售额完成率、人事部门关注的员工离职率、管理人员被考核的 KPI（关键绩效指标）等等，指标是企业管理中核心的、重要的数据资产。

下图清楚地展现了当今的指标报告所存在的问题，如果没有集中的指标平台，指标逻辑将在不同的工具中重复定义，导致指标不一致。



图表 5、指标平台架构（来源：Benn Stancil）

一位数据工程师描述了缺少统一指标平台的痛苦：“每两天都需要手动创建新表，但无法判断是否已经存在类似的表。我们数据仓库的复杂性不断增加，数据的来源和变换过程变得无法追踪。当上游发现并修复数据问题时，无法保证修复会传播到所有下游作业。结果是，数据科学家和工程师花费了大量时间来修复数据差异，到处灭火，还非常郁闷。”

指标平台是上游数据源和下游业务应用程序之间的中间层，它被称为指标平台（Metrics Platform）、无头商业智能（Headless BI）、指标层（Metrics Layer）或指标存储（Metrics Store），都是指同一个东西。

与传统的 BI 报告不同，指标平台将指标定义与 BI 报告和可视化分离。拥有指标的团队可以在指标平台中定义他们的指标，形成单一的事实来源，并能够在 BI、自动化工具、业务工作流以及高级分析中一致地重用这些指标。

2.2.2.8 日志分析

日志分析主要服务于 IT 运维。IT 运维是一项庞杂的系统工程，包括网站的运维、系统的运维、网络的运维、数据库的运维、应用系统的运维、桌面端的运维，以及运维开发、运维安全。

运维工作需要借助监控软件，但由于系统庞杂和需求众多，没有任何一款监控软件能够覆盖所有的运维需求，现在大量的运维团队需要通过日志来进行运维管理。

日志的类型很多，主要包括系统日志、应用程序日志、网络设备日志、数据库日志、安全日志等等。每条日志都记载着时间戳、相关设备名称、系统名称、应用名称、使用者及操作行为等相关的描述，系统运维和开发人员可以通过日志了解软硬件信息、检查配置过程中的错误及错误发生的原因。

随着设备、系统、应用、用户数量的增多，设备 7x24 地持续运行，很快就会产生海量的日志数据，一套基于大数据和人工智能技术

的智能运维体系成为必需。鉴于 IT 运维市场有着庞大的体量，代表着 IT 运维未来的智能运维市场将会有巨大的增长空间。

2.2.3 大数据应用

大数据应用，是建立在大数据基础设施之上，综合运用大数据分析和人工智能工具，结合应用场景和垂直行业需求的应用实践。经过 20 多年的发展，大数据应用已经深入社会的各个领域，水平场景应用涉及的领域有：销售、客户体验/服务、企业市场营销、消费市场营销、人力资本、法律、合规、财务、自动化和机器人流程自动化 RPA、安全、广告等，垂直行业应用涉及的领域有：互联网（电商、社交、生活服务等）、金融（借贷、投资、保险等）、电信、政府、卫生健康、工业、交通、教育、房地产、商务、生命科学、农业等。

大数据应用的真正落地，需要结合每个特定用户的特定需求，不是简单的产品堆砌，要做好与用户既有应用环境的结合，并建立新的业务流程。下表举例说明典型的垂直行业大数据应用：

行业	行业挑战	大数据应用	大数据应用价值
互联网	业务场景复杂，数据来源多；业务快速变化，时效性要求高；数据量巨大但数据价值低。	用户行为分析、转化分析、留存分析、活跃分析、渠道分析、个性化推荐、精准营销、广告投放	提升客户满意度、快速获客/留客、提升收入、指导产品开发/迭代

金融证券	资金成本高，惠普信贷竞争激烈，信用卡欺诈，证券欺诈，超高频交易。	风险分析，隐私计算，交易前决策支持分析，情绪测量，预测分析，交易数据分析	高度依赖大数据进行风险分析，包括反洗钱，企业风险管理，了解客户和减少欺诈
政府	政府数据资产的整合、管理和开放，政府部门及附属机构之间数据的互联互通。	行程大数据辅助防疫，气象大数据服务于救灾，工商企业大数据检测企业异常等	数据多跑路群众少跑腿，更高效的社会化服务，更卓越的营商环境

图表 6、典型的大数据行业应用

2.2.4 大数据开源项目

大数据技术门槛和项目成本的快速下降，开源大数据项目功不可没。至今，这些开源项目也依然是引领大数据技术发展和创新的重要策源地。

领域	开源项目
框架	Hadoop HDFS, Spark, Hadoop MapReduce, Flink, YARN, TEZ, Kubernetes, Apache Kylin, MESOS, Docker, CDAP, RedHat, HELIX
数据格式	ICEBERG, Parquet, Apache Hudi, ORC, Arrow, DELTA LAKE
查询/数据流	Spark SQL, Pig, Hive, Presto, Apache DRILL, SLAMDATA, GraphQL, Trino, Google Cloud Dataflow, HAWQ, Apache Trafodion

数据访问	Uber Databook, Aundsen, Magda, Ckan
数据库	PostgreSQL, MySQL, MongoDB, GreenPlum, Redis, CockroachDB, MariaDB, Influxdb, Presto, Druid, Cassandra, Airbnb Dataportal, SciDB, DataHub, Apache Flume, Cloud Spanner, CouchDB, Riak, OpenTSDB, Apache Accumulo, ClickHouse, Pinot, EdgeDB, Apache HBase
编排	Apache Airflow, Prefect, Dagster, Flyte, MetaFlow, Kedro, Spotify Luigi
基础设施	Apache Zookeeper, Apache Ambari, Apache MESOS, Argo
数据运营	MARQUEZ, Great Expectations, Open Lineage, LakeFS, Project Nessie
流与消息	Spark Streaming, Kafka, beam Pulsar, Flink, Storm, Apex, Apache NiFi, Apache RocketMQ, Samza
统计工具和语言	Python, R, Scala, NumPy, Pandas, SciPy, RStudio, Pyro, Julia, Tidyverse
AI/机器学习	TensorFlow, Torch, Transformers, OpenCV, Apache MADlib, Scikit-learn, Keras, BERT, XGBoost, Caffe, Microsoft Cognitive Toolkit, DMTK, OpenAI, PyTorch Lightning, Theano, PaddlePaddle, Apache Singa, DIMSUM, FeatureFU, VELES, Mxnet, Neon, Chainer, Uber Michelangelo, ONNX, WEKA, Ludwig, CoreNLP, DSSTNE, MLlib, DL4J, Mahout, Aerosolve, fast.ai, MLR, OpenML, MindsDB, spaCy, Kubeflow, AllenNLP, CatBoost
机器学习运营/基础设施	Pachyderm, MLflow, Kubeflow, mleap, DVC, Seldon, Snorkel, Polyaxon, BentoML, MediaPipe
搜索	ElasticSearch, Apache Solr, Apache Lucene, Sphinx, Sonic, MeiliSearch, Toshi Search, Tantivy, Typesense
日志与监控	ElasticSearch, Logstash, Kibana, Sentry, Prometheus, Fluentbit, Fluentd, Grafana, Vector, Open Telemetry

可视化	D3, Superset, matplotlib, Metabase, Redash, TensorBoard, Seaborn, Bokeh, ggplot2
协同	Beake, Jupyter, Zeppelin, Anaconda
安全	Apache Ranger, Knox, Sentry, Apache Accumulo, Snyk

图表 7、开源大数据项目

2.2.5 数据源和数据资源

数据是新时代重要的生产要素，是大数据应用的基础，数据与应用的相互促进推动了大数据产业更快地发展。多维度的数据接入是大数据应用提升效能的根本保证，而应用的丰富则能更快地提升数据的获取和积累。

在增强分析中，实现数据的自动补充和技术准备，维度丰富的数据接入是基础，包括公开领域的媒体信息、社交动态、气象数据、财经数据、统计信息等等，以及需要协议接口的企业信息、人员信息、财税信息、金融信息、信用信息、地图数据、地理信息、天空海洋数据、环境数据等等。

据工业与信息化部 2021 年 11 月发布的《“十四五”大数据产业发展规划》，我国的数据资源极大丰富，总量位居全球前列。这其中，政府拥有大量高质量的数据，这些数据资产的整合和安全地开放，是正在持续开展的重要工作。

2.3 大数据分析的价值

大数据分析是大数据产业的重要组成部分，核心的价值就是从海

量的数据中找出隐藏的模式、相关性和其他规律，为业务决策提供依据。在企业界常见的大数据分析的价值场景包括：

- **客户获取和保留**

  从维度丰富的消费者数据，可以分析出当前消费者的各种不同类型、个性化的特征以及各自的业务潜力，能够更好地理解不同客户的不同需求，使得企业可以制定有针对性的措施获取新客户，以及提升老客户的满意度。

- **精准营销提升营业额**

  从消费者的购物订单、产品浏览历史、页面停留时间等信息，大数据分析可以描绘出不同的用户画像，进而向消费者推送精准的广告信息，与传统的非精准广告相比，精准广告能够大幅提高成交率。

- **产品开发**

  大数据分析可以提供洞察力，指导产品可行性、开发决策、进度评估。大数据分析能力让产品迭代形成闭环，通过不断地对产品指标进行测量和分析，使产品团队能够知道：新功能是否值得做，用户是否喜欢这些功能、新功能是不是反而给用户添乱。

- **供应链和渠道分析**

  预测分析模型可以帮助进行抢先补货、构建智能的供应商网络、库存管理、路线优化和潜在交货延迟通知等等。

- **风险管理**

  大数据分析可以从数据模式中识别出新的风险，从而制定有效的

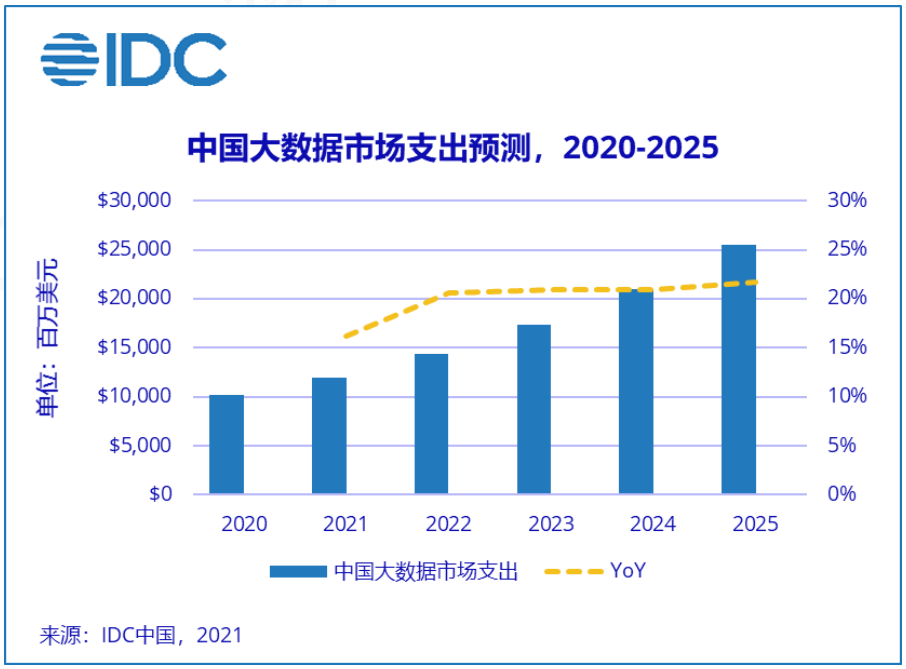
风险管理策略。依靠大数据分析技术构建出来的风控模型，是当前所有小微贷款业务的基石。

3 大数据分析市场规模和发展趋势

自 2014 年以来，“大数据”每年都写入国务院《政府工作报告》，成为国家重点战略。工信部《“十四五”大数据产业发展规划》阐明，大数据产业将保持高速增长，到 2025 年，大数据产业测算规模突破 3 万亿元，年均复合增长率保持在 25%左右，创新力强、附加值高、自主可控的现代化大数据产业体系基本形成。

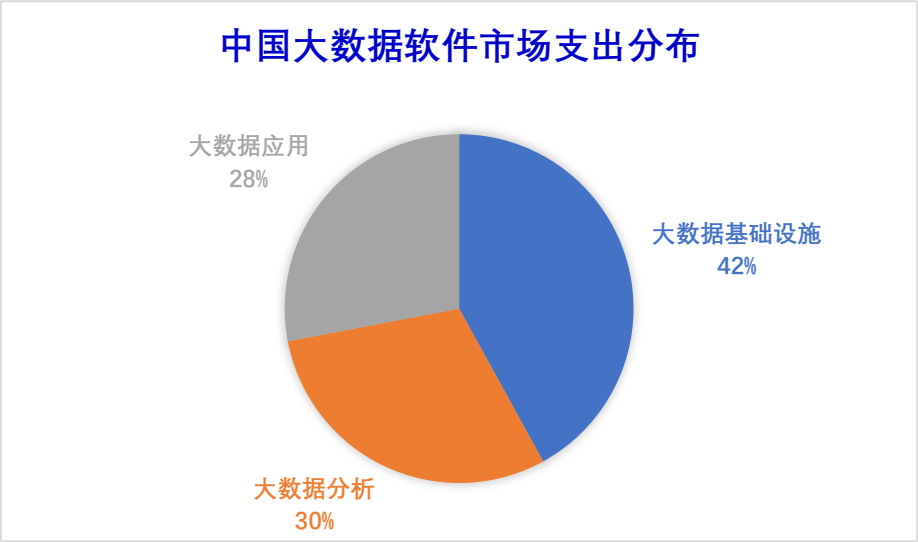
3.1 大数据分析市场规模

据 IDC 预测，中国大数据市场 2021 年整体规模超 110 亿美元，且有望在 2025 年超过 250 亿美元，呈现出强劲的增长态势。



图表 8、中国大数据市场支出预测 2021v2（来源：IDC）

至 2025 年，预计中国大数据硬件市场约占 40%，超过软件和服务，增长稳定；大数据软件市场占比将逐年提升，2025 年超 30%的市场支出将流向软件。



图表 9、中国大数据软件市场支出分布（来源：中国大数据网）

中国大数据网对大数据软件市场的进一步细分做了独立研究，2021 年中国大数据软件市场支出中，大数据基础设施占比为 42%，大数据分析占比为 30%、大数据应用占比为 28%。以此推算，2021 年中国大数据分析市场支出为 10 亿美元，2025 年有望超过 22 亿美元。

3.2 大数据分析市场趋势

大数据分析市场具有如下趋势：

- 国产化产品蓬勃发展
- 云化部署持续增长，公有云、非公有云部署同步发展
- 大数据分析平民化

3.2.1 国产化产品蓬勃发展

国产化的大数据基础实施及大数据分析产品蓬勃发展，相较于国外产品在细分领域的专精，国产化产品的发展更多地体现了集成，这也体现了国内用户的需求特点，即需要覆盖数据整合、数据加工、数据治理、数据分析、数据大屏的全链条需求，国产化产品更能适应这种市场需求。

3.2.2 云化部署持续增长，公有云、非公有云部署同步发展

大数据时代的数据源不仅仅是数据库时代的结构化数据，而是存在越来越多的半结构化和非结构化数据，它们几乎是天然地放在云计算环境中。国内云环境的实际部署与国外较大规模使用公有云的状况不同，中国政府用户和一些重点行业（如金融、电信等）对公有云的使用还是相对较少，这些客户更多地使用自己的私有云/行业云环境，因此中国用户的大数据硬件采购金额占比也高于国外，本地部署及私有云/行业云模式仍需要采购大量硬件设备，大数据软件也需要部署到属地环境中。

3.2.3 大数据分析平民化

与大型企业不同，数量巨大的中小企业更愿意采用公有云的方式进入大数据领域，云计算大数据平台开箱即用、按需付费的模式极大地降低了中小企业使用大数据的技术门槛和资金投入门槛。云计算大数据平台背后的专业技术公司，能够将技术进步的成果第一时间提供

给用户，人工智能技术带来的效率提升以及技术门槛的降低，可以快速普惠到最终用户，大数据分析开始平民化，大量的业务人员开始从事大数据分析工作，而无需专门的技术人员支持。反过来，用户数量的大幅度增加也进一步促进了数据分析技术和业务的发展，形成良好的生态。

3.3 大数据分析技术趋势

大数据分析的技术发展趋势如下，它们都与人工智能和架构统一相关：

- 增强分析步入人工智能阶段
- 湖仓一体成为新的数据基础设施底座
- 流批一体将两种架构模式融为一体

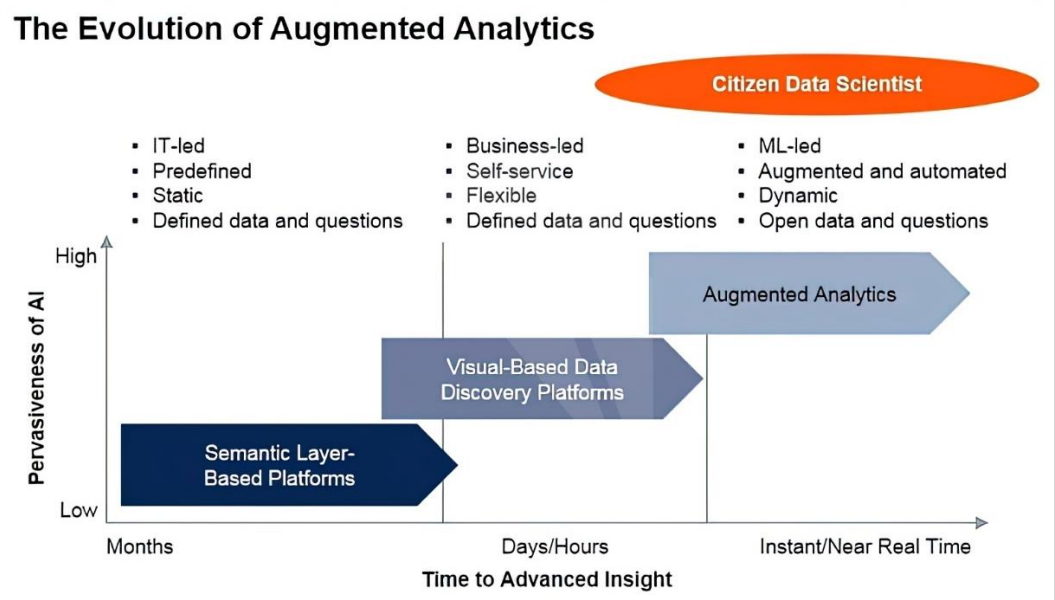
3.3.1 增强分析步入人工智能阶段

增强分析演进到了人工智能时代，使用机器学习来自动化的数据准备、洞察发现、数据科学以及机器学习模型开发、业务发现共享，适用于广泛的业务用户、运营人员和数据科学家。它将成为未来大数据分析的核心特性，使数据分析、学习的时间大大降低，可以真正普惠所有的业务用户。在大数据分析的输入、分析、输出三步骤，它都可以大显身手。

输入步骤中，增强分析可以帮助用户从杂乱无章的数据源中找出最有价值的部分，推荐或者按照用户意愿补充外部数据（例如做区域

销售额分析的时候补充当地的人口和 GDP 数据），并做好数据类型、数据格式转换等耗时费力的技术性准备工作。

在分析步骤中，增强分析可以帮助用户做自动建模、模型管理、代码生成，操作的技术门槛大大降低，使用户可以把更多精力放在业务发现本身而不是繁琐的技术细节。机器学习所擅长的，正是从海量数据中发现规则，也是人类难于胜任、易于忽略、效率低下的领域。



图表 10、增强分析的演进（来源：Gartner）

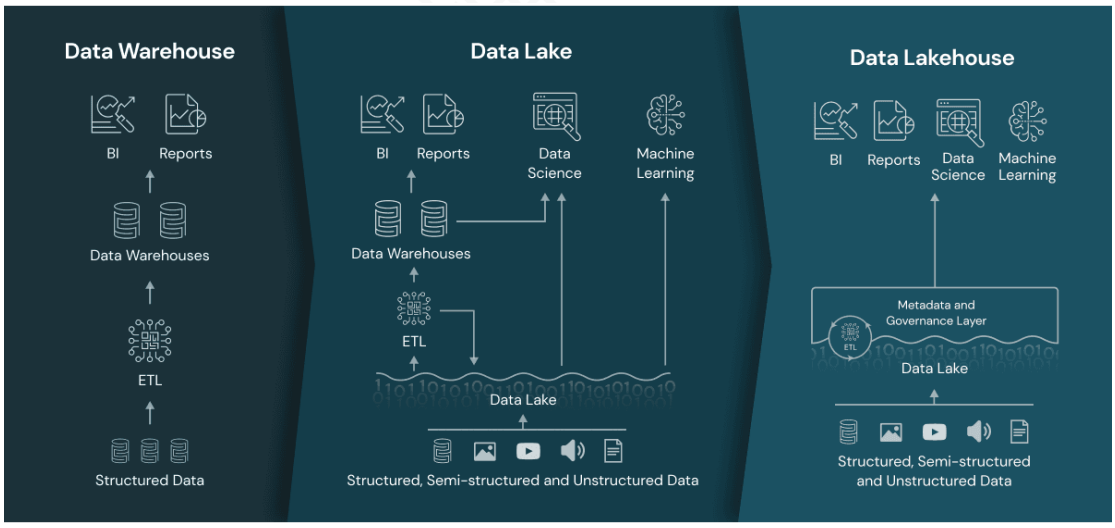
在输出步骤中，增强分析给出的不仅仅是一堆柱状图、饼图、仪表盘，用户还需要在其中摸索才能找到业务发现。增强分析能够自动地提出业务发现，用于更广泛人群的交流，包括业务人员、运维人员和数据科学家。

增强分析在商业智能、智能运维等大体量市场有着广阔的发展空间，拥有坚实的 AI 技术并且在合适的业务场景中辛勤耕耘的厂商将获得快速成长的机会。

3.3.2 湖仓一体成为新的数据基础设施底座

数据仓库与数据湖和核心差异在于存储的不同。数据仓库源自数据库，而数据湖源自云存储上的开放格式数据文件，因为应用需求不同、发展路径不同，形成了两套不同的技术栈。而随着应用的不断深入发展，企业和互联网应用需求走向融合，对数据仓库和数据湖的两套技术栈也就自然提出了融合要求。湖仓一体是大数据基础设施的未来，是数据分析所依赖的基础平台。

当前为了满足不同的需求，大型用户需要同时建立数据仓库和数据湖两套系统，带来数据重复、数据一致性差、存储成本高、数据报告和数据分析应用结果不一致、数据缺乏治理时效性差、技术不兼容等一系列问题。



图表 11、数据仓库、数据湖、湖仓一体架构对比（来源：databricks.com）

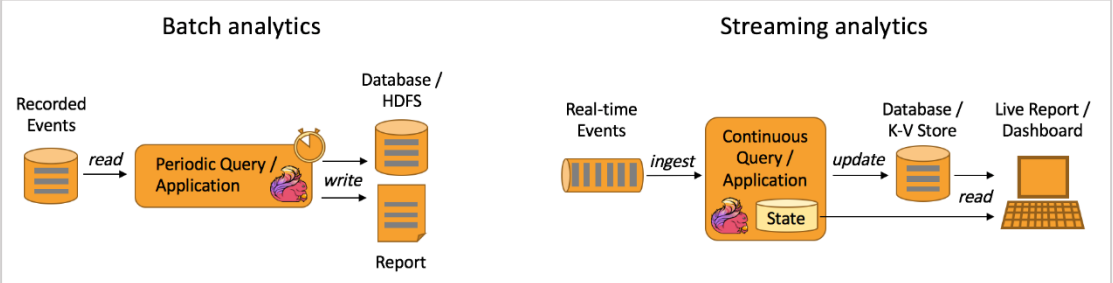
湖仓一体是对数据仓库和数据湖的融合发展，支持数据湖所欠缺的原子化事务、元数据目录、高性能，也具有数据仓库所欠缺的数据治理能力（数据仓库里迷宫一样的 ETL 很常见，到处留下不同的数据

副本，变更管理繁重而复杂），对接庞大的开放式软件生态系统（数据仓库只能用 SQL），最重要的是支持机器学习和人工智能（数据仓库不支持 ML/AI 使用的稀疏数据集，例如视频、音频、任意文本）。

在数据分析领域，湖仓一体是未来，也是全流程流批一体化的基础。它可以更好地应对 AI 时代数据分析的需求，在存储形态、计算引擎、数据处理和分析、开放性以及面向 AI 的演进等方面，要领先于过去的分析型数据库。

3.3.3 流批一体将两种架构模式融为一体

流批一体是指将流式计算与批量计算两种不同架构的数据处理模式融合到一起，对混合的有界数据和无界数据能够统一进行支持，提供更一致的、更广泛的编程环境，以较少资源浪费，降低维护成本，获得更好的数据一致性。



图表 12、批量分析与流式分析（来源：flink.apache.org）

在大数据分析领域，流批一体化的重要意义在于，不管输入的是有界数据（批量处理）还是无界数据（流式处理），可以采用同一套查询接口，并产生同样的结果。而分别采用两套系统时，需要不同的查询系统，产生的结果也不一定一致。

4 大数据分析三大细分市场主要厂商分析

如“图表 2、大数据产品分类”所示，大数据分析产品可细分为：商业智能平台、数据可视化、数据分析师平台、增强分析、数据目录与发现、指标平台、流批一体化、日志分析、查询引擎、搜索等细分领域。在中国市场，市场份额最高的是商业智能 BI 市场，数据可视化与 BI 密切相关，数据分析师平台、增强分析也主要应用于 BI 领域，所以我们归入同一个市场；数据目录与发现、指标平台、流批一体化属统一架构之列，我们归入流批一体化市场；日志分析、查询引擎、搜索的重要应用场景是智能运维市场。本行业研究主要对商业智能和数据可视化、流批一体、智能运维三个市场展开讨论。

中国大数据网通过对全国工商企业数据进行挖掘分析，截至 2021 年 12 月底，中国大数据企业共约有 6.53 万家。中国大数据网以六大维度（行业实力、身份特征、创新能力、活跃程度、发展速度和经营风险）为指标，对上述约 6.53 万家大数据企业进行了量化分析，并按照指标高低，划分三级九等（AAA、AA、A、BBB、BB、B、CCC、CC、C，由高到低排列），实现了对大数据企业的综合科技创新能力的量化评价。中国大数据网对其中 A 级和 B 级的企业进行了入库，并提供系统平台以供社会各界查询和应用。“中国大数据网认证企业查询系统”入口为：http://cxpj.handsdata.net。

从大数据企业的运营情况来看，整个大数据分析市场仍然遵从

20/80 规律，即整个行业的大部分市场份额、主营业务收入、利润等仍然被属于少量的主要企业掌握，大部分企业仍在艰苦努力寻求发展。本行业研究报告主要针对上述大数据分析市场的主要厂商进行分析。这些主要厂商的运营情况基本代表整个产业发展的现状和趋势。

在海量数据支撑下，中国大数据网将大数据分析细分市场内的主要厂商分为成熟型厂商和新兴型厂商两类，他们各自有不同的特点。

厂商类型	特点
成熟型厂商	为体量较大的互联网巨头、大型 IT 服务商、大数据概念上市公司等，技术、市场实力雄厚，产品线一般比较广泛、完整。
新兴型厂商	多为新兴创业公司，勇于创新，一般专注于特定领域，在某些前沿技术上处于相对领先位置。

图表 13、大数据分析市场厂商类型

成熟型厂商基本为全球知名企业、上市公司或者科技头部企业，拥有强大的品牌号召力、行业影响力和营收能力，绝大多数已经处于发展的中后阶段，业务呈现多元化态势。新兴型厂商多为新兴科技企业，绝大部分坚持聚焦在细分领域，勇于创新，开拓进取，不断在细分领域取得优秀成绩，甚至处在细分市场的技术领先位置。新兴型厂商是科技创新最活跃的群体之一，拥有广阔的发展空间。

本行业研究报告将研究重点主要聚焦在新兴型厂商，以更好地反应细分市场的科技创新活跃程度，积极鼓励和支持中小型科技企业的发展。

下表是大数据分析整体市场的主要厂商，入选的依据是：在大数

据分析领域有较高的技术声誉、市场声誉，在中国大数据网对大数据分析厂商调研分析中获得专家团队的认可，并且 2020 年主营业务收入排名靠前。下表以厂商中文名称的拼音为序。

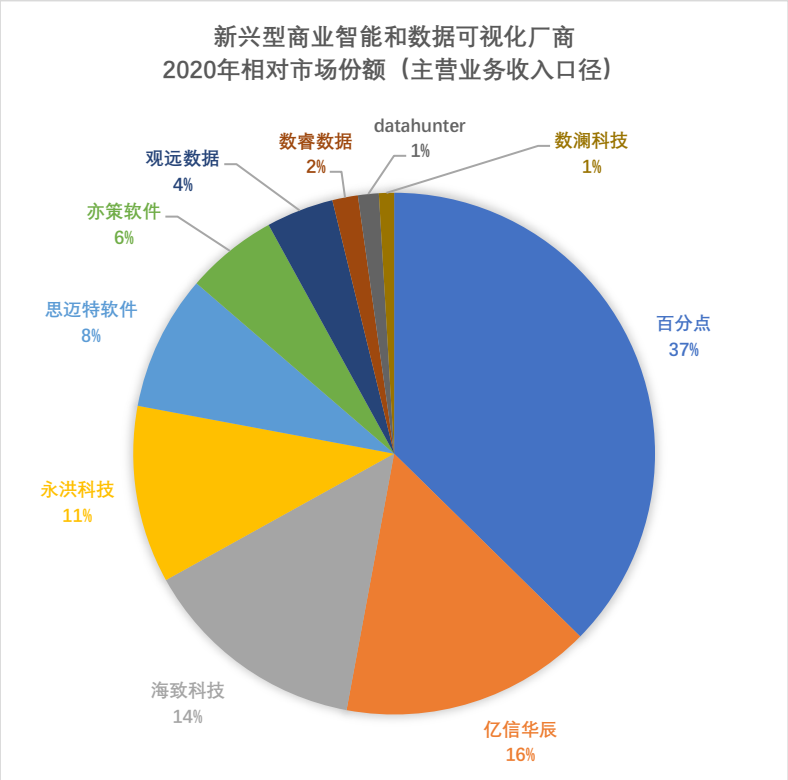
图表 14、大数据分析市场主要厂商

4.1 商业智能和数据可视化

商业智能和数据可视化是大数据分析中最大的市场，主要厂商如下：

图表 15、商业智能和数据可视化市场主要厂商

入选的新兴型厂商，商业智能和数据可视化都是其主营业务之一，下图是这些厂商 2020 年主营业务收入口径的相对市场份额分布情况。



图表 16、新兴型行业智能化和数据可视化厂商 2020 年相对市场份额（主营业务收入口径）

（来源：中国大数据网）

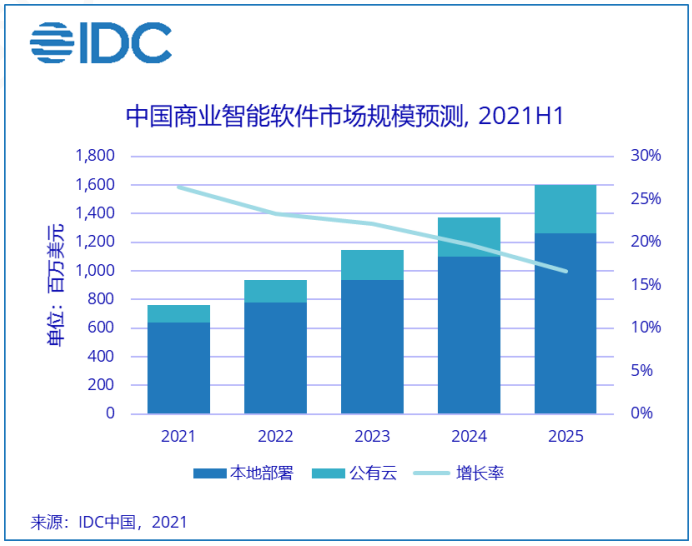
中国大数据网对上述新兴型厂商同时进行了综合科技创新能力评价，评价结果如下表。（备注：综合科技创新能力评价是基于中国大数据网的积累的大数据资源以及分析模型输出的结果，可能与其他研究机构采用的评价标准和使出结论存在差异，仅供参考。）

序号	厂商	评价等级	综合评价	综合评分
1	数澜科技	AAA	非常健康	865
2	数睿科技	AAA	非常健康	864
3	观远数据	AAA	非常健康	858
4	思迈特	AAA	非常健康	847
5	百分点	AAA	非常健康	845
6	亿信华辰	AAA	非常健康	823
7	永洪科技	AAA	非常健康	798
8	海致科技	AAA	非常健康	798
9	亦策软件	AAA	非常健康	797
10	DATAHUNTER	AAA	非常健康	761

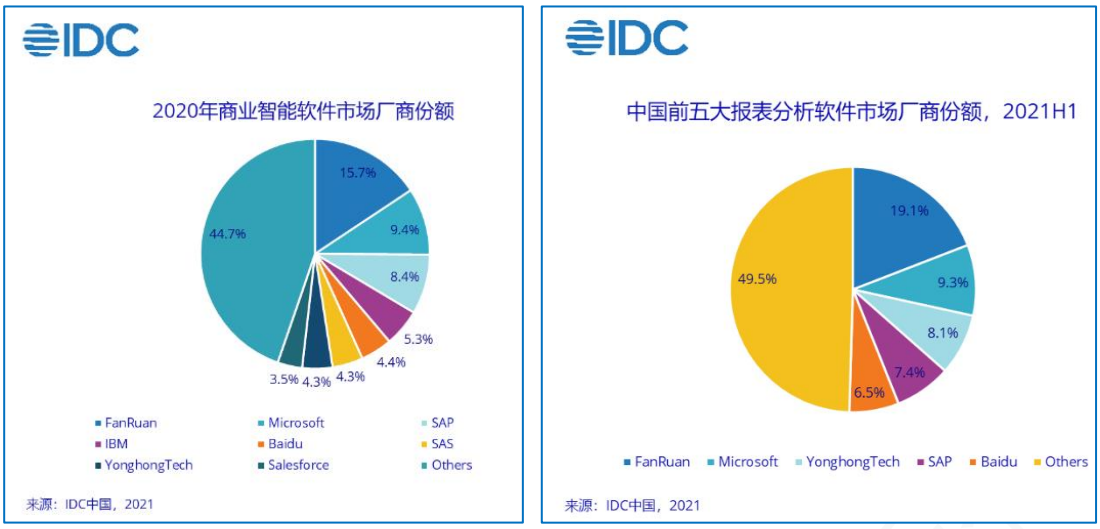
图表 17、新兴型行业智能化和数据可视化厂商综合科技创新能力评价

（来源：中国大数据网，截至 2022 年 4 月 20 日）

商业智能和数据可视化是大数据分析市场中规模占比最大的，据 IDC 预测 2025 年中国市场可达到 16 亿美元，占大数据分析市场的 70%以上。



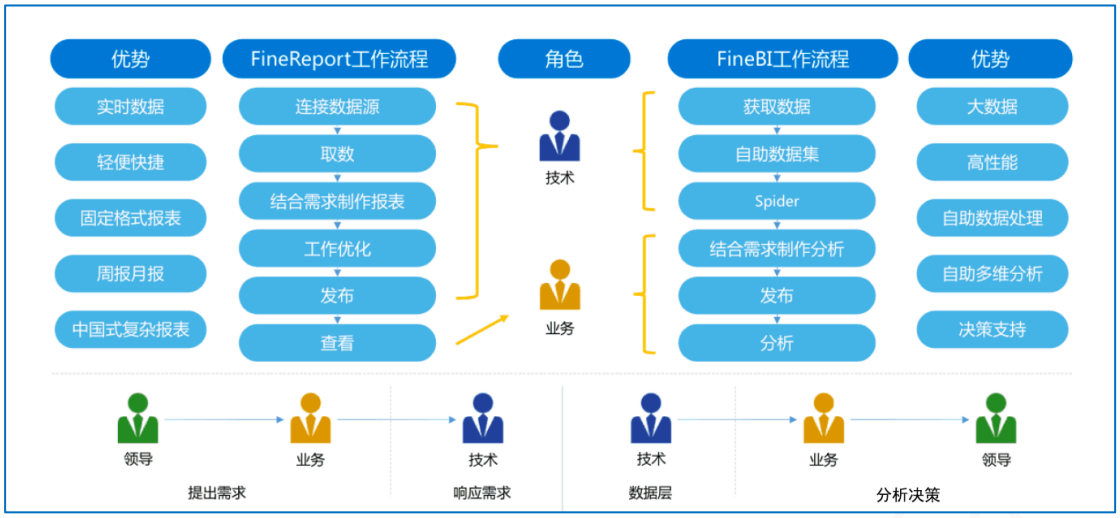
图表 18、中国商业智能软件市场规模（来源：IDC）



图表 19、中国商业智能和数据可视化软件市场厂商份额（来源：IDC）

IDC 数据中，2020 年商业智能软件与 2021 年报表分析软件两次市场份额统计的口径有差别，一次偏向商业智能，一次偏向数据可视化，两者有区别又有关联，可以看到两次统计中的前两名厂商都是帆软和微软。

帆软是中国商业智能领域最早的厂商之一，长期专注耕耘商业智能市场，2021 年以 200 多人的团队规模实现超 11 亿元销售收入，累计合作客户数超过 18000 家，中国 500 强企业中有超过 300 家是帆软的客户。微软则是全球的软件巨头，在商业智能领域更是全球的领导者之一，有着丰富的开发技术生态和合作伙伴网络。这里面有个有趣的现象，就是在国际上处于商业智能领导地位的 Tableau 没有进入到中国市场份额的前列，这有 Tableau 产品对中国用户需求适应的问题，也有市场经营策略的问题。



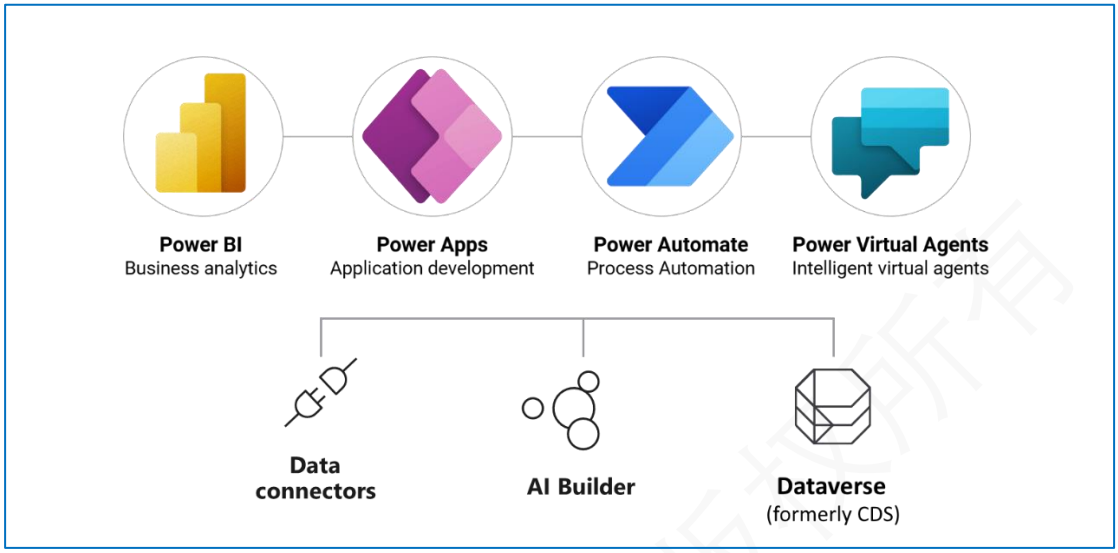
图表 20、帆软的商业智能产品

从几十年前的数字报表到现在的可视化展现，商业智能的市场需求是巨大的，通常它的用户是销售人员、管理人员、财务人员、以及专门的数据分析师，而 IT 技术人员反而不是商业智能的主流用户。非 IT 技术人员作为主要用户，他们的需求是容易生产、容易消费。容易生产，是指用户能够快速地创建一份新的数据报告，最好能够自己动手操作，技术门槛低、易学易用，以前的数据报告总需要找技术人员开发，业务与技术之间沟通非常费劲而且开发周期漫长；容易消费，是指整个流程使用顺畅，能够随时随地访问，支持手机端。这两点正是数据分析平民化最核心需求。

帆软和微软的商业智能产品作为中国市场上的领导者，都顺应了数据分析平民化的需求，主要手段是：云化和低代码。

云化以及其上的手机端支持，极大地提升了产品的访问便利和使用流程的顺畅。2014 年，帆软推出了独立 SaaS 产品“简道云”，先从中小企业服务开始，与钉钉的合作，让简道云能够无缝融入到中小企业的日常的工作流程之中，非常符合中国用户的使用习惯。微软的

Power BI 则秉承着微软的整体云化战略，可在公有云或混合云 + 本地方案中快速部署，并提供丰富的移动端支持。



图表 21、**微软的** Power Platform

在低代码方面，帆软的简道云以钉钉为渠道，为用户提供低代码开发工具，近几年发展迅速，已服务全国超过 100 万家大中小型用户。而微软则以 Power Platform 为更大的平台，除了商业智能组件 Power BI 之外，还包括 Power Apps、Power Automate、Power Virtual Agents 等组件，可与微软 Office 365、Dynamics 365、Azure 以及第三方应用程序无缝集成，提升企业快速构建解决方案的能力。

至于近年火爆的 AI 技术，帆软的态度相对稳重，在产品实现上采取了稳扎稳打的策略。

微软作为人工智能的巨头之一，AI 本就是其长期的战略方向，自然有更多的成果可以放入商业智能产品，目前 Power BI 可以支持语音交互，也具有多种 AI 分析模板可供选择。从整个 Power Platform 的产品结构上来看，更强悍的 AI 功能组件是与 Power BI 商业智能组

件相独立的，具有更好的可扩展性，用户可以综合考虑自身的业务需求和性价比，做出最适合自己的产品组合选择。

相对于帆软相对稳健的 AI 策略，新兴型厂商则更积极地拥抱“BI+AI”。永洪科技的 Z-Suite AI 深度分析平台，百分点的 Clever BI，观远数据的智能 ETL 和 AI 模型实验室等产品，数睿数据的增强分析型敏捷商业智能平台 Nextion BI，海致科技的 AI+BI 产品生态，思迈特软件的 SmartBI Cloud 增强分析云平台，这些产品及配套服务都是新兴型厂商谋求在“BI+AI”新赛道上获得突破的努力。

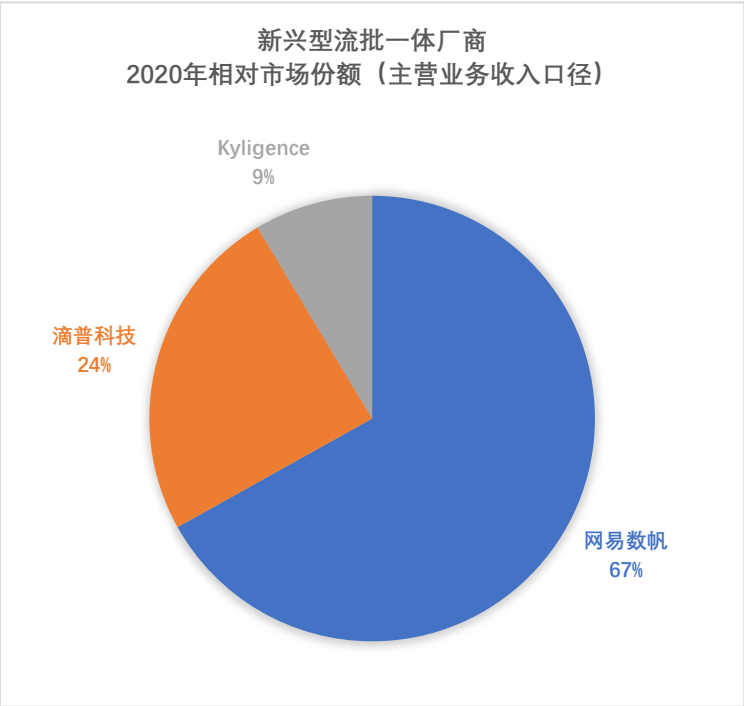
4.2 流批一体

流批一体分为存储、计算、分析应用三部分，存储的部分——湖仓一体市场启动最早，几乎所有的大数据基础设施厂商都在跟进，计算和分析应用部分目前处于相对早期阶段，赛道上的厂商还相对较少，未来发展有较大空间。这个领域的主要厂商如下。

图表 22、流批一体市场主要厂商

入选的新兴型厂商，流批一体都是其主营业务之一，下图是这些

厂商 2020 年主营业务收入口径的相对市场份额。这三家新兴型厂商基本代表了阿里等云计算巨头之外的流批一体细分市场，无论是在技术创新，还是在业务探索上都为细分市场的发展做出了重要贡献。



图表 23、新兴型流批一体厂商 2020 年相对市场份额分布（主营业务收入口径）

（来源：中国大数据网）

中国大数据网对上述新兴型厂商同时进行了综合科技创新能力评价，评价结果如下表。（备注：综合科技创新能力评价是基于中国大数据网的积累的大数据资源以及分析模型输出的结果，可能与其他研究机构采用的评价标准和使出结论存在差异，仅供参考。）

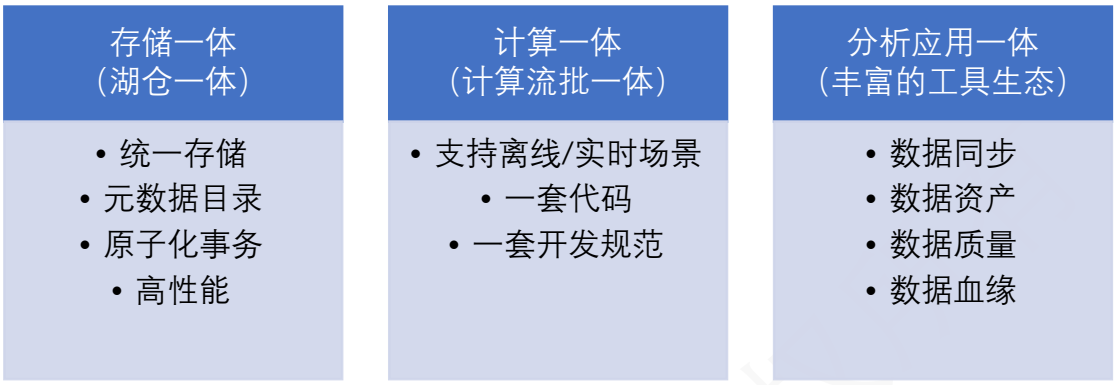
序号	厂商	评价等级	综合评价	综合评分
1	网易数帆	AAA	非常健康	827
2	Kyligence	AAA	非常健康	791
3	滴普科技	AAA	非常健康	780

图表 24、新兴型行业流批一体化厂商综合科技创新能力评价

（来源：中国大数据网，截至 2022 年 4 月 20 日）

流批一体，是指离线处理、实时处理全流程的一体化，需要覆盖

三个部分：存储一体（湖仓一体）、计算一体（即计算流批一体）、分析应用一体（丰富的工具生态）。我们认为广义的流批一体包含以上三个部分，而狭义的流批一体是其中的计算部分。



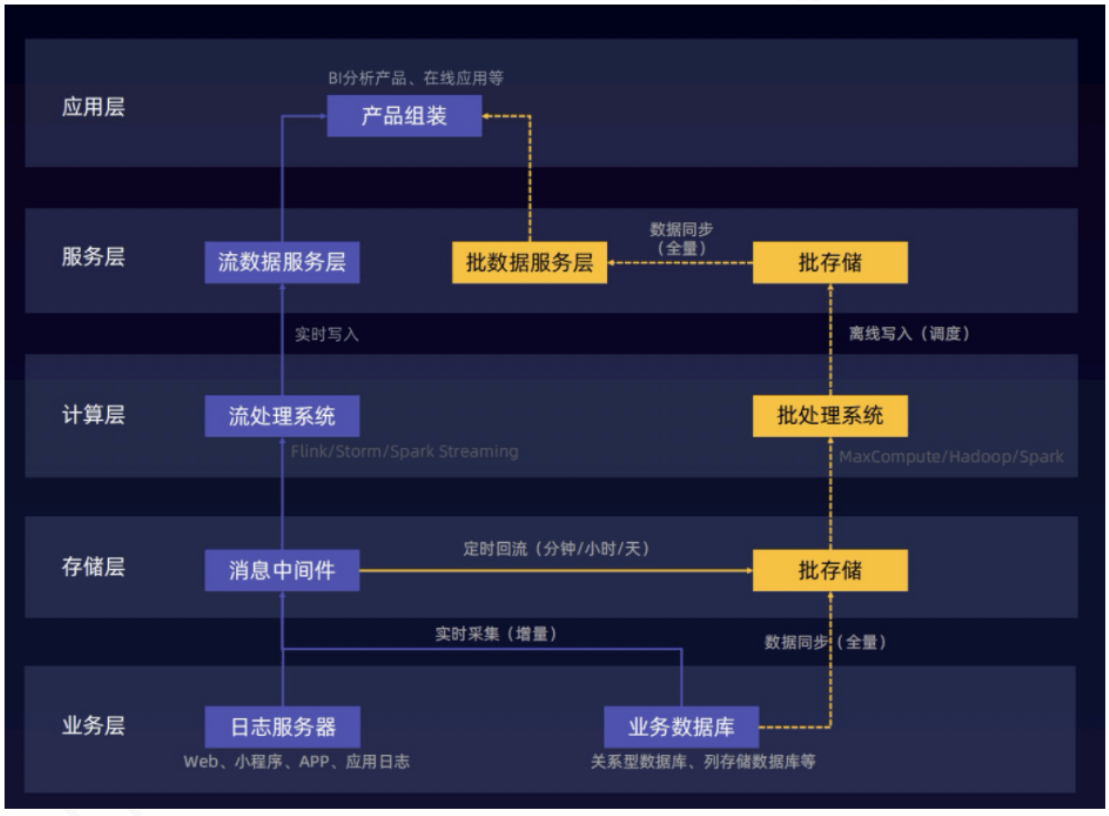
图表 25、广义流批一体的三个板块

存储上的湖仓一体属于大数据基础设施领域，作为一个新兴的数据架构，湖仓一体正成为兵家必争之地。国际上的 DataBrick、Snowflake 是领导者，AWS、阿里云、华为云等头部云厂商提供云化湖仓一体支持，国内的上百家数据库/数据仓库厂商也都在跟进。对聚焦存储一体的厂商我们在此不做展开，而是重点讨论具有广义流批一体能力的厂商。

流批一体领域，阿里无疑是最活跃的厂商。2019 年初阿里巴巴集团斥资 9000 万欧元收购了 Apache Flink 框架背后的德国公司 Data Artisans，阿里巴巴本身也是 Flink 最大的用户之一，Flink 开源社区正是计算流批一体最活跃的贡献者和推动者。从 Apache 软件基金会 2020 年度报告可以看出，在反映开源社区繁荣情况的三个关键指标上 Flink 都名列前茅：用户邮件列表活跃度，Flink 排名第一；开发者提交次数 Flink 排名第二，Github 用户访问量排名第二。这些

数据并不局限于大数据领域，而是 Apache 开源基金会下属的所有项目。

阿里的双 11 大促是每年的业务重头戏，大促期曾经面临离线（批量处理）和实时数据（流式处理）统计口径不一致的问题，“秒秒钟几百万上下”的数据差异，会影响广告、商务甚至公司运营决策，强电商属性和大业务体量倒逼着流批一体技术必须在阿里核心业务落地，方能解决痛点。



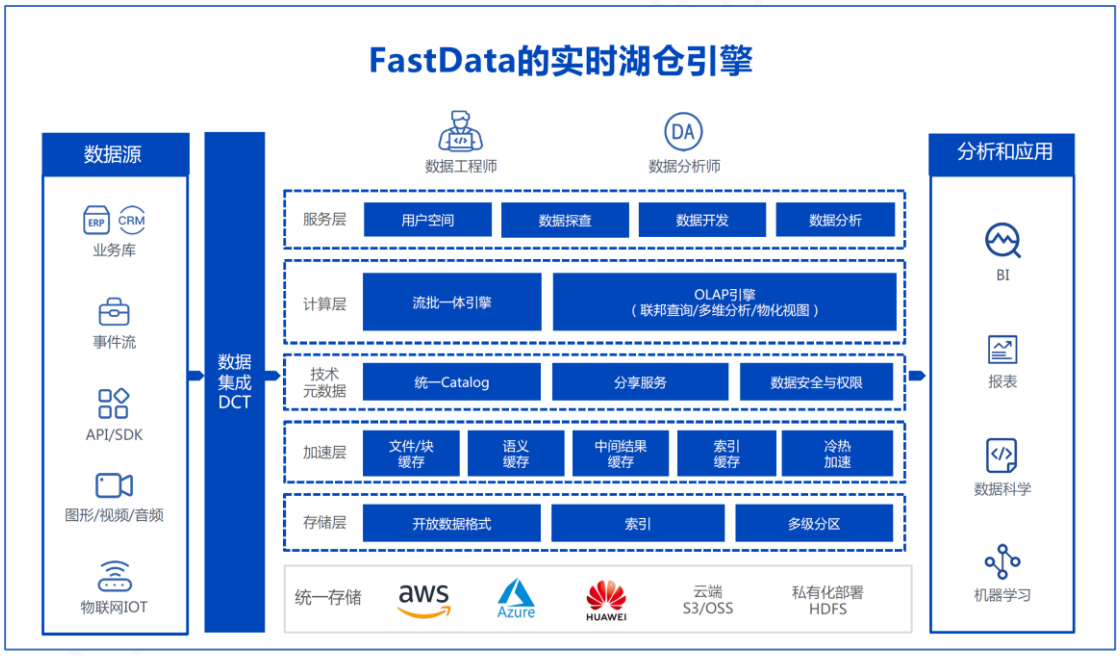
图表 26、阿里的流批一体架构

2020 年双 11，基于 Flink 的流批一体应用出现在阿里最核心的数据业务场景，并抗住了 40 亿条/秒的实时计算峰值，这是流批一体技术真正规模化落地应用于超大规模核心数据业务，这也意味着 Flink 在阿里的发展已经进入第二个阶段，从全链路实时化进阶到全

链路流批一体化。

除了阿里之外，其他的成熟型厂商也在跟进流批一体，华为 Data Lake Insight、新华三 DataEngine 等都包含有基于 Flink 的流批一体的产品和解决方案，而腾讯则入资了流批一体的初创公司。

在新兴型厂商中，流批一体作为一个相对新兴的领域，形成产品和解决方案的厂商目前还不多，其中综合排名靠前的厂商是滴普科技。滴普科技 FastData 的实时湖仓引擎已经在一些行业用户中落地应用，包括半导体工厂、车联网用户、时尚运动集团、电气制造集团、药业公司等。

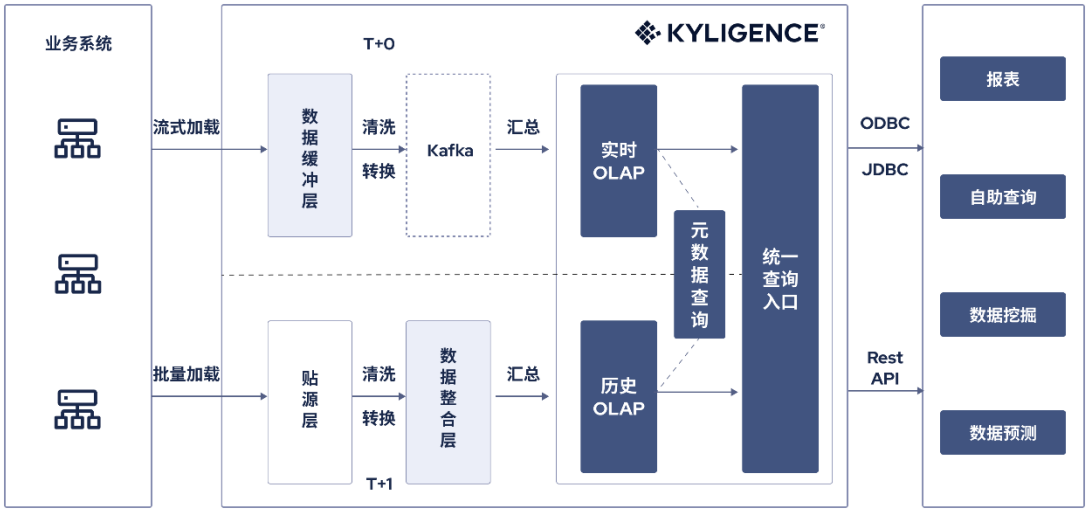


图表 27、滴普科技 FastData 的实时湖仓引擎

滴普科技的 FastData 云原生数据智能平台，具有湖仓一体、批流一体的架构，体现了存储/计算/分析应用这三方面的构成，有较全面的组件以适用于不同的场景需求。滴普科技的理念是保持产品的云中立和互联网生态中立，使其产品能够在任何云、任何互联网生态中

都有成长的土壤。

Kyligence 则将流批一体作为一个解决方案为客户提供服务：在数据处理阶段，通过同一套计算框架来处理历史和实时数据；在数据分析阶段，向用户提供历史数据和实时数据的融合查询能力。仅通过一个数据模型、一个 SQL 语句，通过统一的查询出口，同时接入批数据和流数据，实现数据分析的流批一体。Kyligence 的流批一体方案已在银行业获得落地用户。



图表 28、Kyligence 的流批一体解决方案

网易数帆则认为广义流批一体的三个方面都是网易大数据平台未来的重点改进方向。即做到大数据平台的实时化，而不是将实时计算独立出来做。针对存储的流批一体，现在已经有实时数据湖引擎 Arctic，适配了 Flink、Spark 计算引擎，Impala、Presto 的适配也在进行中。同时，开发流批一体和工具流批一体方面的工作也在紧锣密鼓地展开。

4.3 智能运维

智能运维，是基于运维基础信息和运维日志数据（各种系统日志、监控信息、应用信息等），运用人工智能和机器学习来提升运维自动化程度的运维方式，它是大数据与 AI 技术相结合的综合应用。从市场分类上而言，智能运维是大数据分析与大数据应用相结合的交叉领域市场。

智能运维是 IT 运维的未来，随着 IT 系统的体量越来越大、复杂度越来越高，原来依赖于人力和传统 IT 服务管理工具的运维方式无以为继，必须借助 AI 和机器学习的能力大幅度提升运维效率。2020 年中国 IT 运维服务行业市场规模超过 2500 亿元，年复合增长率虽然减缓，但依然保持在 10%以上（来源：华经产业研究院），市场空间巨大。目前国内智能运维的上市企业新炬网络（605398.SH）2021 年收入为 5.9 亿元、博睿数据（688229.SH）2021 年收入 3.6 亿元（估），只是大运维市场的一个很小的零头。智能运维尚处于起步阶段，产品成熟度处于初期，市场也相对分散，还有巨大的成长空间。这个领域目前的主要厂商如下：

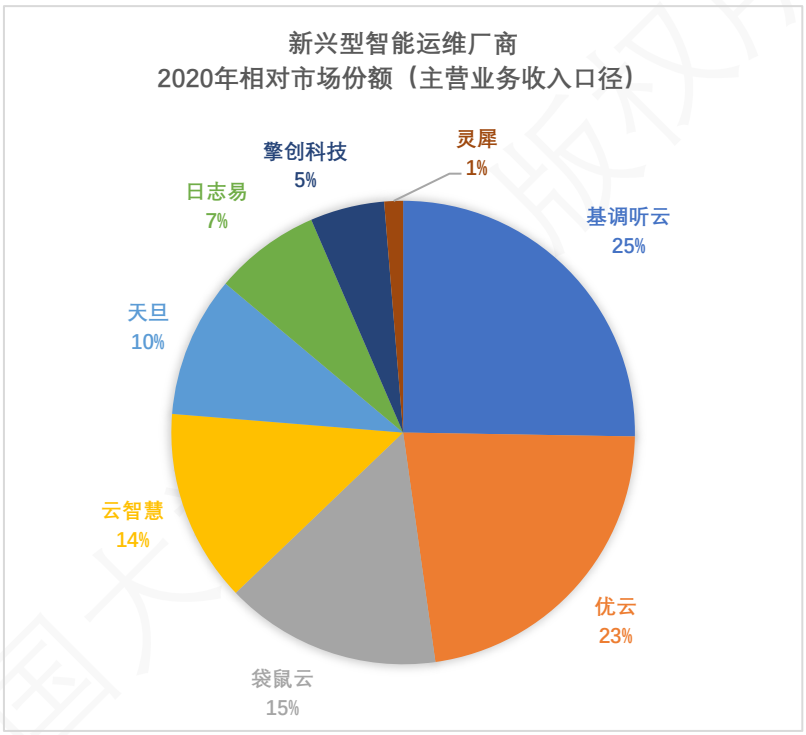
智能运维领域主要的成熟型厂商

splunk> 新炬网络 SNC Net Bonree 博睿数据 | 股票代码:688229 aws 阿里云

HUAWEI Baidu 百度 腾讯云 Microsoft AsiaInfo 亚信科技

图表 29、智能运维市场主要厂商

入选的新兴型厂商，智能运维都是其主营业务之一，下图是这些厂商 2020 年主营业务收入口径的相对市场份额。



图表 30、新兴型智能运维厂商 2020 年相对市场份额分布（主营业务收入口径）

（来源：中国大数据网）

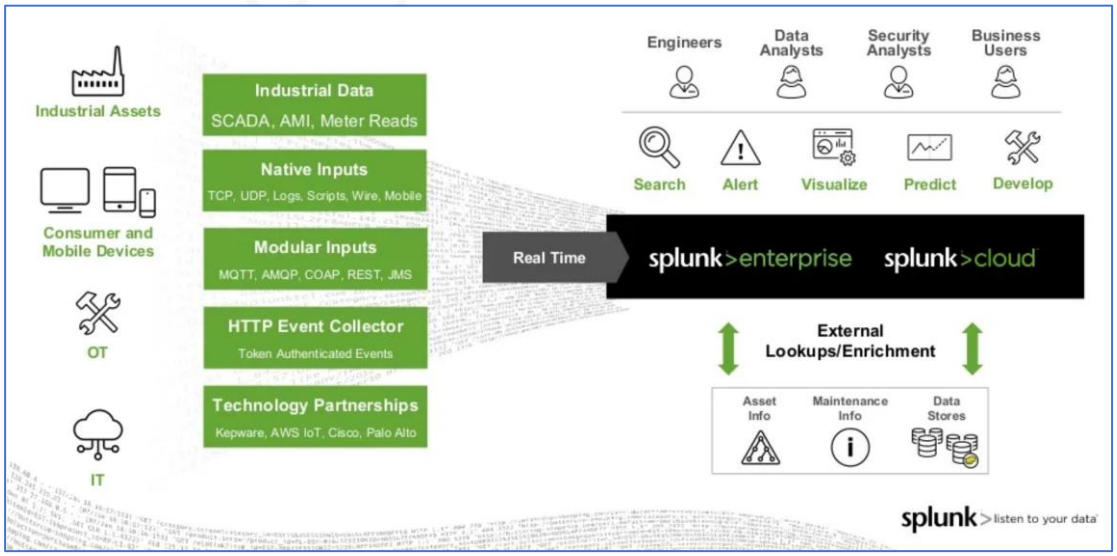
中国大数据网对上述新兴型厂商同时进行了综合科技创新能力评价，评价结果如下表。（备注：综合科技创新能力评价是基于中国大数据网的积累的大数据资源以及分析模型输出的结果，可能与其他研究机构采用的评价标准和使出结论存在差异，仅供参考。）

序号	厂商	评价等级	综合评价	综合评分
1	天旦网络	AAA	非常健康	862
2	日志易	AAA	非常健康	861
3	优云	AAA	非常健康	848
4	基调听云	AAA	非常健康	847
5	擎创科技	AAA	非常健康	823
6	灵犀	AAA	非常健康	776
7	云智慧	AAA	非常健康	760
8	袋鼠云	AAA	非常健康	734

图表 31、新兴型智能运维厂商综合科技创新能力评价

（来源：中国大数据网，截至 2022 年 4 月 20 日）

在成熟型厂商中，包括亚马逊云、阿里云、百度云、华为云、腾讯云、微软云在内的各大云服务厂商，在运维信息及系统日志上的数据基础都很坚实，可以说这是公有云最基础的功能之一，这些信息作为数据源就可以作为输入送入各种分析处理系统之中，可以是云服务商的标准分析系统，用户也可以进行进一步的开发及与其他系统集成。云厂商提供的运维支持在其云内是统一有效的，但对于云外系统的支持则不那么有效。



图表 32、Splunk 智能运维平台

全球范围来看，Splunk 是智能运维领域的领先厂商之一，能够实时接收来自各种数据源的运维数据，对数据进行搜索、报警处理，并在此过程中运用机器学习能力，获得 IT 服务智能、进行用户行为习惯分析，Splunk 提供丰富的全链条工具。

图表 33、新炬网络的全栈一体化智能运维平台

作为国内上市公司和智能运维的龙头企业，新炬网络通过多云全栈的“服务+产品”模式持续打磨并推广智能运维产品，补充云服务商的运维短板。客户拓展方面，在电信、金融、交通运输、先进制造等各行业的新客户和新项目拓展上发力；产品研发方面，形成了智能运维全系列产品体系，基于自身运维实践并运用自身的专家系统和运维场景，打造了基于产品中台的新炬网络智能运维全系列产品和运维数字员工产品。

图表 34、博睿数据智能运维监控产品

同属成熟型厂商的博睿数据是发展迅猛的一家，于 2020 年 8 月科创板上司，是中国应用性能管理 APM 技术的领导厂商，同时具有数据分析/AI（人工智能）能力。依托完整的 IT 运维监控能力，公司利用大数据和机器学习技术构建的先进智能运维监控能力，可基于自身的通用性，满足最为广泛的用例，有效控制企业成本，确保数字化业务平稳运行，保证成功交易，保障良好的数字化体验，更有针对性地向客户提供服务。

图表 35、基调听云智能运维产品

基调听云是新兴型厂商中收入体量位居前列的厂商，推出了覆盖 IT 运维领域的一系列产品，通过人工智能技术在运维领域的应用与

创新，帮助企业简化复杂的 IT 运维世界，智能发现深藏的问题，加速业务创新实现，最终实现人与系统的智能化高效协作。

在新兴型厂商中，擎创科技在智能运维领域活跃度很高，是被 Gartner 连续推荐的国内智能运维 AIOps 领域标杆供应商，专注于将人工智能赋能运维管理，激活运维数据智慧，助力客户数字化转型。

图表 36、擎创科技智能运维平台

擎创夏洛克 AIOps 智慧运营平台，以全局运营视角解读 IT 运维，整合告警、性能指标、日志等多维数据，在 AI 算法运维中台的支撑下实现精准告警、异常检测、根因定位和容量分析等智能场景，助力企业降本增效、优化运营决策。擎创科技目前已在银行、金融、制造、能源交通等多个行业拥有成功案例。

智能运维有着广阔市场前景，新兴型公司很多，活跃的公司还有：云智慧、灵犀、优云、日志易、天旦、袋鼠云等等。

5 结论

工信部《“十四五”大数据产业发展规划》指出，“十四五”时期是我国工业经济向数字经济迈进的关键时期，对大数据产业发展提出了新的要求，产业将步入集成创新、快速发展、深度应用、结构优化的新阶段。《规划》部署了 6 项重点任务：一是加快培育数据要素市场，二是发挥大数据特性优势，三是夯实产业发展基础，四是构建稳定高效产业链，五是打造繁荣有序产业生态，六是筑牢数据安全保障防线。在自主可控和开放合作的大时代背景，以及《规划》的部署指导之下，我国大数据分析领域还将会有长足的发展。

总结前文，我们对未来的大数据分析市场有如下观点：

- 大数据基础设施中的湖仓一体化已经成为此领域厂商的必争之地，也是大数据产业赖以生存的未来基础，“十四五”期间国产化数据库、数据湖、数据仓库、湖仓一体化产品将有长足发展。未来的大数据分析将以湖仓一体为基础设施。
- 流批一体化也正成为新的市场机会，一体化架构会有更多的大型客户采用，并成为未来新的架构标准，被更多中小型客户效仿。
- 人工智能/机器学习正在与传统大数据应用快速结合，在商业智能以及所有大数据应用领域都展现出了强大的生命力，人工智能算法与大数据应用场景相结合中蕴藏着庞大的市场机会。
- 基于日志大数据和人工智能技术的智能运维是传统 IT 运维的未来，市场体量巨大，有望快速成长出更多大体量的公司。

6 研究机构简介

中国大数据网是为贯彻落实国家大数据发展战略，促进我国大数据与科技传播应用的发展，推动大数据与科技传播应用人才的成长，聚合产业资源，助力科技信息更高效传播而组建的经国家相关部门核验通过，由中华人民共和国工业和信息化部备案的（京 ICP 备14029848 号-5）大型科技信息类网站，是中国科协直属国家一级学会中国科技新闻学会主办的大数据与科技传播专业委员会官方工作平台。中国大数据网名称使用具备合法性和唯一性。中国大数据网是中国科技新闻学会大数据与科技传播专业委员会的官方工作平台。中国大数据网的企业主体平台为中科新数字（北京）科技信息有限责任公司。

中国大数据网栏目内容涵盖了大数据、云计算、人工智能、物联网、5G、区块链、元宇宙等领域，按照“平台+智库+产业”的模式，打造聚合“产学研用”资源的核心枢纽；通过智库体系支撑，形成推进产业科技创新服务的核心价值；为产业数字化转型升级，构建联通“政企产学研”和提升品牌影响力的直通桥梁。中国大数据网将根据产业发展需要推出多样、细化的区域和行业发展指数，并持续组织行业交流活动，帮助各个城市汇集企业、人才、技术、资本等资源，不断推动产业科技创新活动向纵深发展。

中国科技新闻学会是中国科协直属的由中央人民广播电台、中央电视台、人民日报社、新华社、光明日报社、经济日报社、科技日报

社等单位共同发起，1988 年 1 月正式成立的全国性一级法人社会团体，也是世界科技记者联盟的发起者和成员之一。大数据与科技传播专业委员会于 2017 年 4 月根据民政部民发[2014]38 号和中国科协办公厅科协办函学字[2016]246 号文件规定，经中国科技新闻学会第五届八次常务理事会议审议通过并成立，是中国科技新闻学会在大数据与科技传播应用领域的专业机构。工作方向涉及科技新闻、数据新闻、新媒体、科技报刊、科技影视、大数据传播、智能传播、科学普及、媒体工作、组织教育、国际交流、太空文化、品牌传播、农业科技、科技创新、健康生态等相关领域。主办有《中国科技信息》、《新媒体研究》、《科学家》、《电子竞技》、《科学中国人》、《科技创新与品牌》、《科幻画报》、《科技传播》和《中国市场技术报》等权威刊物。

中国大数据网

网址：http://www.zgdsj.org.cn
电话：010-68599082
邮件：info@zgdsj.org.cn
地址：北京市西城区复兴门内大街 45 号院(学会办公区)