企业数据能力测评：认清现状，布局未来丨建设数据中台系列（一）

**“我的企业目前在数据应用上处于什么水平？接下来应该向哪个方向努力？”本文试图帮助企业决策者和 IT 负责人解答这一问题。**

今天，数据之于企业的重要性，已经勿须多言，建设数据驱动型企业也已成为众多企业的战略目标之一。在这一趋势的引领下，很多企业开始了新一代数据平台（例如数据中台）的建设工作，然而在启动这一具有挑战性的工作之前，企业首先需要冷静客观地审视一下自己的数据生态，弄清楚目前所处的能力水平，以及下一步努力的“方向”。只有这样才确保后续工作是沿着正确方向展开的，这可能也是企业在构建全新的大数据平台或数据中台前最先需要弄清楚的问题了。本文核心观点援引自作者所著的[《大数据平台架构与原型实现：数据中台建设实战》](https://item.jd.com/12677623.html" \t "_blank)一书。

1. 如何度量企业的数据应用能力？

企业的数据应用能力决定了企业在“数据”这座金矿中所能攫取的价值大小，既然是一种能力，就会有强有弱，有高有低，收集并统一存储数据只是建立良好数据生态的第一步，数据背后的真正“价值”是需要通过专业的手段进行挖掘才能获取的。我们常说：“如果数据是燃料，那么分析就是引擎”，对于一家企业而言既要储备燃料，也要装配引擎，只有同时具备了数据和数据分析能力才能从数据中提炼出有价值的信息。为了清晰地度量企业在数据应用上的能力水平，我们对数据应用涉及的多个方面进行了归纳和总结，得到一个“企业数据应用成熟度模型”：

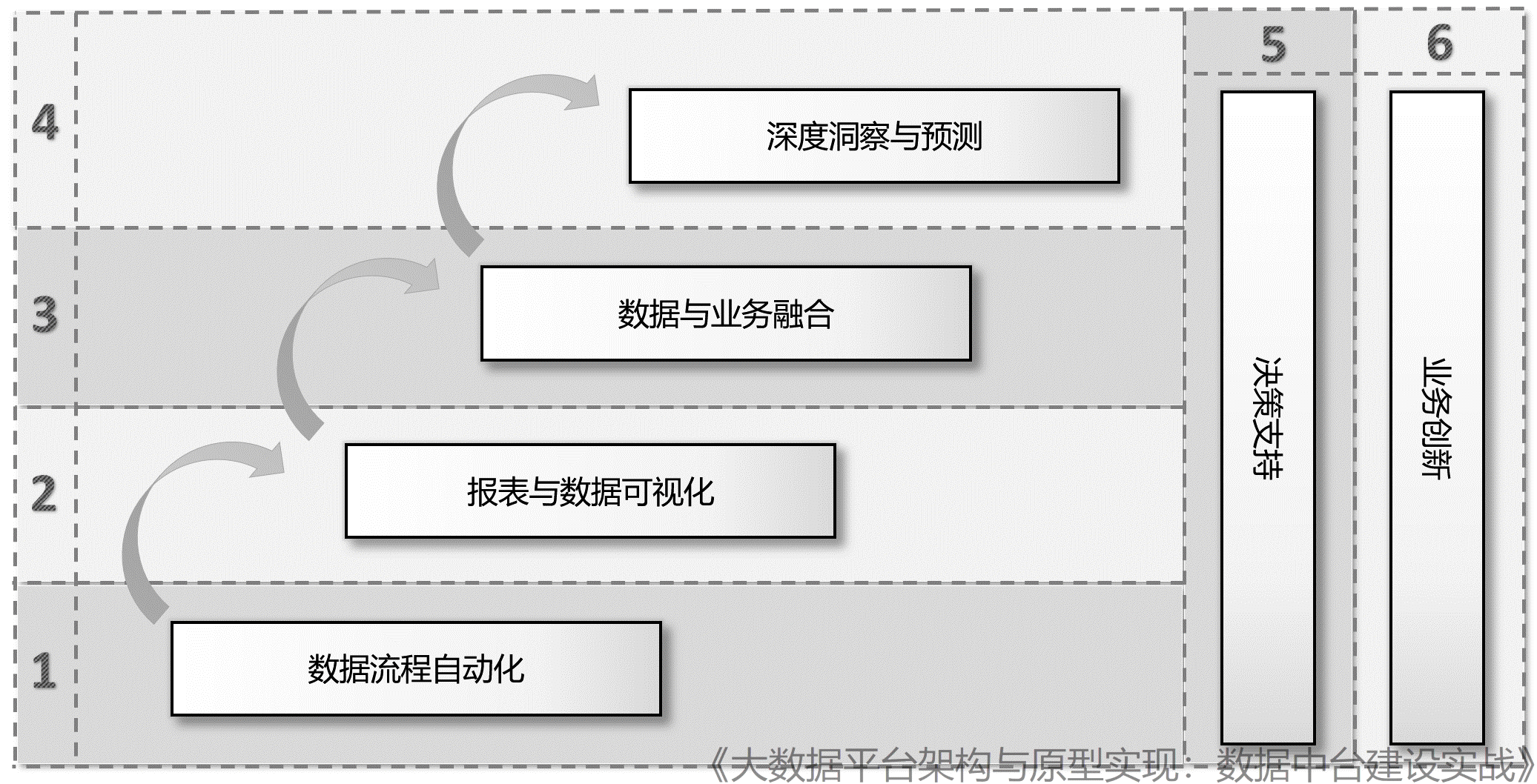


图 1 企业数据应用能力成熟度模型

在这个模型中，我们引入四个等级和两个维度来度量企业的数据应用能力：

1.1. 第一层级：数据流程自动化

数据流程自动化指的是数据从产生的源头到使用的末端是自动化的，中间没有手工操作，全部通过系统集成完成对接。可能有的读者会认为这一能力不应该成为一个独立的等级，因为在高度信息化的企业中应该已经实现了各个系统间的数据对接，即使是以最原始的文件形式交换数据也大都已经实现了流程自动化，然而在很多企业中，实际的情况却并非如大家想象中那样理想。

现实中企业的数据来源丰富多样，既有自身业务系统产生的数据，也有外部系统和供应商提供的数据，还有业务人员日常手工维护的大量表格和纯文本数据，很多企业可能已经完成了对自有应用系统的自动化数据采集与处理，但是对于大量的外部数据和业务人员手工维护的数据往往是还没有建立起有效的自动化处理流程的，这些数往往有这样一些特点：

* 格式不规范
* 经常变动
* 缺乏基本的校验，容易出现错误数据
* 数据供给周期不固定

这些原因导致了这类数据很难被自动化获取和处理，而很多时候这些数据恰恰又是业务流程闭环中重要的组成部分，缺失这些数据会导致无法进行数据分析或极大影响结果的准确。造成这类数据大量存在的原因有两点：

* 企业的信息化程度依然不够，在某些局部业务范围内出现了系统空白，从而需要业务人员手工介入，以文件和表格的方式维护数据
* 企业的数据资产意识不足，对数据规范化的重视程度不够，缺乏一些管控和约束手段。

相应的，企业实现高度的数据流程自动化需要做好如下几点：

* 持续进行企业信息化改造和升级，将 IT 系统覆盖到企业的全部业务流程中，这会很大程度上消除维护手工数据的情况，因为当所有的业务流程都通过 IT 系统来驱动时数据就会沉淀到系统的后台数据库中，且这些数据都已经过了系统的校验和规范化处理，质量都非常高，也可以很方便地从中提取出来放入数据平台
* 从企业管理层开始建立“数据资产”意识，成立专门的数据治理组织，有计划地规范和治理企业的数据生态，对于重要的数据要制定统一规范的格式，避免任意地对数据格式进行改动

1.2. 第二层级：报表与数据可视化

在收集到足够多的企业数据后，就可以开展常规报表和数据可视化的开发工作了，这是目前多数传统企业所处的阶段，它们通过传统的数据仓库技术收集并整理了大部分的企业数据，通过报表工具向业务和管理人员提供一些常规的报表，这些报表通常面向生产、供应链、销售、市场、财务等不同的业务环节，在时间粒度上最细可达 daily 级别。数据的展示的形式多以表格为主，同时也会借助报表工具展示图形化报表。过去的报表大多在 PC 端展示，随着移动应用的兴起，开始出现越来越多面向企业开发的手机 APP 和微信小程序，在这些终端上为业务用户提供报表服务正越来越受欢迎。在这一层级上的企业对于数据处理和分析表现出如下一些特征：

* 基本上完成了与各个业务的系统的对接，数据能被自动化采集
* 已经建立了数据仓库体系，企业数据可以被有效地统一管理
* 已经开发了业务上迫切需要的一些核心报表，业务对数据系统的依赖度高
* 依托于成熟的后台数仓，新的报表和数据展示需求都可以较快地完成开发并投入使用

第二层级是很多企业目前停留的阶段，并且可能在这一层级上停留了很多年，因为很多企业都在这一层级上遇到了“瓶颈”，很难再发展到下一层级，主要原因有以下三点：

* 传统的单体数仓系统缺乏水平伸缩的能力，已经无力应对企业数据爆炸式的增长，不得不放弃和展缓了集成某些新业务数据的计划
* 传统数仓只能处理关系型数据，对于越来越多的图片、视频和其他非关系型数据无能为力，而这些数据往往是由新业务形态产生的，对于这类数据处理能力的缺失会让企业错失新的市场机遇
* 传统数仓只能进行批处理，缺乏实时数据处理能力

如果企业想突破这些瓶颈，就需要将企业数据平台升级为以大数据和 AI 为技术核心的新一代数据平台，然后重建数据版图。

1.3. 第三层级：数据与业务融合

在第二层级时，对于数据的应用只局限在“描述”业务上，而并没有使数据参与到业务中，各种报表在业务用户的工作中扮演的是一种辅助性角色，对于业务的影响是通过业务用户和管理者在数据的支持之下做出的判断和决策，从数据应用的角度看，这是一种被动和滞后的方式，并没有充分的发挥出数据蕴含的潜能。在进入第三层级之后，这个状况会逐渐被扭转，数据开始与业务进行融合，数据以及数据处理能力会全面参与到业务流程的各个环节中去，从而产生更大的价值。

这是一个全新的阶段，是数据驱动型企业在具备了大数据处理能力之后，借助 AI 和机器学习而实现的一种更加智能的企业信息化水平，在这一层级上企业将具备如下能力：

* 数据直接赋能业务，数据分析的结果将直接反馈回业务系统，作为业务系统的某些关键操作的输入
* 已将多种维度的丰富数据进行融合，可以更加准确地刻画数据背后的“事实”
* 已具备实时的数据处理能力，可以让业务用户实时掌握数据
* 大数据平台已经成熟且稳定
* 已经出现基于传统的机器学习和数据挖掘的应用，在某些局部领域已经出现小范围的深度学习案例

第三层级看上去有些抽象，我们可以通过一些案例来理解。例如，客户会员体系是 CRM 系统中非常核心的一个功能，其中的会员积分计算是一个逻辑复杂且计算量又大的操作，消费者的每一笔交易和若干重要的行为数据都会触发积分的计算，传统的 CRM 系统很难实现用户积分的实时计算和更新，而是按天进行批量处理，这样一来，用户体验就会变差。现在很多新的 CRM 系统都在积极地引入大数据的流式计算实时处理用户交易和行为数据，并更新用户积分，这是数据与业务融合的一个非常好的案例，即借助大数据的计算能力来实现业务上的数据处理需求。

另一个案例是用户画像系统，用户画像是基于用户的基本信息，消费记录，社交行为等多种数据进行数据建模之后，利用算法生成的关于用户的一套标签体系，这些标签全面刻画了用户的特征和属性，因此称之为“用户画像”。用户画像在 CRM、精准营销和以用户为中心的产品与服务创新上起着重要作用，是很多 2C 端企业非常重要的一类系统，同时它也是典型的大数据系统，但功能和定位又是业务性极强的系统。

从第二层跃升到第三层时，企业的数据基础设施会面临一次脱胎换骨的革新，传统的关系数据库，数据仓库和 BI 等基础设施已经不能支撑第三层级的诸多需求了，这时需要企业构建下一代的数据平台。业界对于“下一代数据平台”的认知经历过一些更迭，早期方案是使用大数据技术替换传统的数仓系统，后来出现了 datalake——数据湖的理念，其方案还是以大数据作为主要的技术支撑，但是在理念上比传统数仓又有新的创新，而现在业界特别是国内最认可也是呼声最高的方案则非“数据中台”莫属了。

1.4. 第四层级：深度洞察与预测

现在人们的一个共识是：数据除了可以告诉我们现在，还可以“预知未来”，深度洞察与预测是数据金字塔最顶端的价值输出，也是目前我们认为的企业可以达到的最高层级的数据应用能力，即运用 AI 和深度学习算法对数据进行深度地洞察，揭示传统分析方法无法发现的数据特征，并基于现有数据对未来趋势进行预测。

我们来看一下企业到达第四层级后会具备哪些能力。前面我们提到的智能门店选址的案例就是第四层级上的一个代表案例，对于零售行业来说，门店选址是非常重要的，会直接影响到零售商的销售业绩，传统选址的做法是通过人工现场勘查，然后经过主观判断做出决定，这种方式选出的门店其实际效果难以量化，成功率无法确定，而如果能够基于人口、消费、竞争对手、环境业态和交通路网等丰富的多维度数据再配置适当的人工智能算法进行综合分析是可以得出更加精准的选址方案的，并且不单单是门店位置，还可以给出门店的预计销售额、门店产品的上货策略等更加细致和完备的数据。

另一个示例是智能客服系统，这类系统可以针对顾客提出的问题进行语义识别，然后根据提出的问题在知识图谱中进行搜索，寻找匹配的答案或决策，人工智能客服可以 7\*24 小时在线，随时解答顾客的问题，既提高了客户满意度又能节省商家的人力成本。

以上四个层级并不一定非要自下而上逐层构建，实际上很多企业的数据生态是在上层业务的驱动下自然形成的，并不会像模型中描述的这样层次分明，但是能力模型能给到企业管理者一个清晰的认识：即自身企业目前整体上停留在哪个层级上以及接下来应该向哪个方向发展。

1.5. 两个纵深维度：决策支持与业务创新

最后，在成熟度模型图的右侧，还有两个贯穿始终的维度：决策支持与业务创新，它们即是企业构建数据平台进行数据分析的价值导向，也是企业数据应用能力持续输出的效果，企业达到的层级越高，对于决策支持与业务创新起到的作用就越大越明显。

在这两个纵深维度上，企业更多需要做的是建立业务与技术团队的互信，以产出业务价值，进行业务创新作为共同的追求，同时要潜移默化地培育企业的“数据文化”，在企业内部形成“用数据说话”的氛围。

所以这两个纵深维度更多考察的是企业在数据方向上的管理、协作以及企业文化，是一个需要从管理者开始营造并推动，企业全员参与的良性互动过程。在这一过程中，业务团队需要培养更好的数据素养，善于通过数据分析业务现状，并辅助自己作出决策，而技术团队不断加强数据平台的各项能力，确保更好的服务于业务分析，同时积极主动地学习业务，从 IT 视角为业务创新提供新鲜素材。

2. 如何度量企业的数据技术水平？

前面我们是从数据应用的“效果”上观察企业的数据能力，当落地到实现层面时，“技术”就是不可或缺的了，构建数据平台通常是从基础设施建设开始的，然后配合业务上的需求，逐步完善和打通各个技术环节，在这里，我们不讨论传统技术框架下的构建路线和方案，因为正如我们在上一节中提到的，如果企业想晋升到第三或更高的层级，就需要以大数据技术作为基石构建新的数据平台，所以我们下文讨论的所有技术内容都是以大数据作为背景展开的：

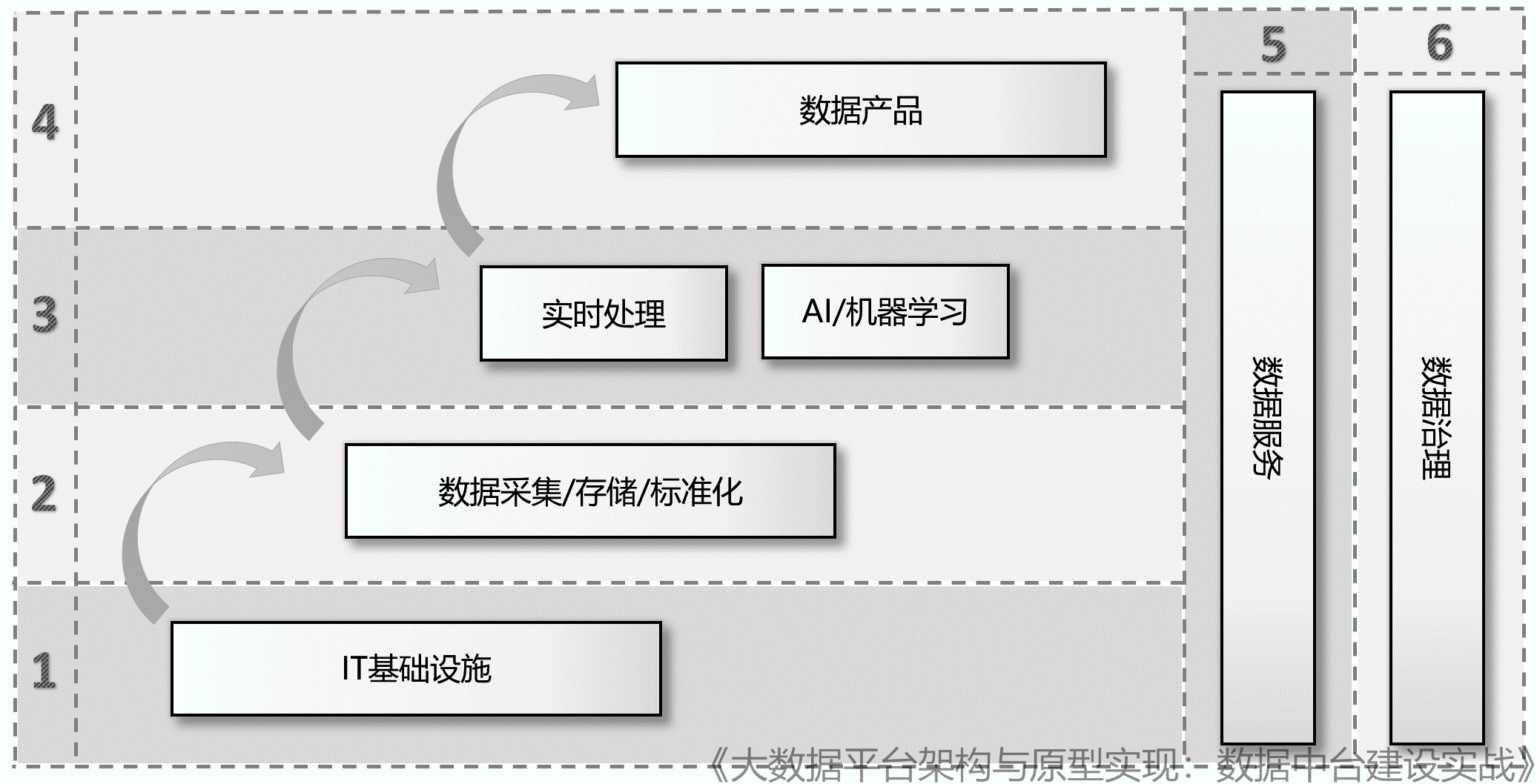


图 2 企业数据技术成熟度模型

2.1. 第一层级：IT 基础设施

首先，IT 基础设施是当然的前置条件，构建基础设施包括硬件机器的安装，组网和调试，操作系统和必要软件工具的安装，然后，在硬件资源之上安装和维护一个大数据集群，这个集群将负责承载企业全部数据的存数和处理任务。如果再宽泛一些，用于支撑平台运行的基础服务，例如 DevOps，数据和算法服务使用的容器和容器编排服务等也都算在基础设施内。

过去，企业的 IT 基础设施大多是通过自有机房或租赁数据中心的硬件设施完成，随着云计算的普及，越来越多的企业开始把部分的基础设施迁移到云平台上，形成混合的云架构。基于云平台的基础设施在运维的便捷性、系统的可伸缩性和成本控制上都有显著的优势，同时，在 PaaS 层面上，云计算厂商也提供对标 on-premise 的数据平台服务，如 AWS 的 EMR 等，这些因素促使促使越来越多的企业将新一代数据平台建设在云基础设施之一。但是是在云平台上企业需要特别重视数据安全问题。

2.2. 第二层级：数据采集 / 存储 / 标准化

有了必要的基础设施之后，就可以展开数据的采集、存储和标准化工作了，这一工作也可以简单地表述为数仓的建设过程。这一阶段需要将分布在各个业务系统里的数据收集起来，在进行一些必要的规范化处理之后，存储在一个统一的大数据平台上，这是一个长期的迭代过程，特别是在建设初期，上层对数据的广泛需求和下层集成数据源的繁重工作之间会存在冲突，我们建议企业通过启动一到两个大型项目来驱动这一阶段的建设工作，然后在中后期维持一个规模较小的团队持续跟进其他数据源的接入工作，当企业在这一层级积累一段时间后，就可以交付相应的报表和数据可视化应用了。

2.3. 第三层级：实时处理、AI/ 机器学习

再接下来，进入第三层级就要将技术平台推升到更高水平了，这里有两项非常重要的技术拓展：实时处理和 AI/ 机器学习，这是现代大数据平台两项标志性的技术能力。实时处理是指通过流式计算、NoSQL 数据库等技术实现大体量数据的实时处理和读写，实时的数据处理能力对一些事实性要求很高的业务场景至关重要，这也是以往传统数据平台很难做到的。由于实时处理对技术和研发人员的要求都更高，因此大多数企业一般会先完善平台的批量处理能力，然后再逐步拓展到实时处理领域。

另一个领域就是 AI/ 机器学习方面的建设了，这一领域对技术能力的要求更高，且参与人员的角色和背景也与传统的 IT 人员有所不同，进入到该阶段时，IT 团队需要引入数据科学家，算法工程师等 AI 领域的人才。最后，实时处理和 AI/ 机器学习这两大领域的能力是可以同步培养的，彼此之间没有太大的依存关系。当企业具备了第三层级的技术能力之后就可以有力的支撑应用能力模型中的最高层级“深度洞察与预测”了。

2.4. 第四层级：数据产品

最后，从纯技术的角度上还有一段可以上涨的空间，那就是以业务领域为划分依据，将现有各个层级上的技术能力进行提炼并培育成“数据产品”，从功能、性能、灵活性和可扩展性等多种维度上进一步提升数据平台的技术成熟度。

2.5. 两个纵深维度：数据服务和数据治理

与四个层级建设并行的还有两项贯穿始终的工作：数据服务和数据治理。数据服务是指将数据平台上的各种数据以服务的方式提供给其他系统，这种“服务”可以通过 Restful API，JDBC，ODBC，FTP 等形式或协议体现出来，这是将数据应用能力辐射到企业的各个系统与业务领域上的关键一步，没有灵活而有效的数据接口，数据平台在企业范围内起到的作用会受到限制。与此同时，数据治理也是一个长期的持续性的工作，数据治理就是对企业的数据资产进行清晰的梳理，明确管理职责，建立配套的标准规范，同时要确保所有策略和规范能落地执行，数据治理的最终目的就是保障数据质量。

应用能力成熟度模型和技术成熟模型之间是有关联的，根据我们的经验，当企业的技术成熟度达到第二层级时，可以支撑应用能力成熟度的第二层级和部分的第三层级，当技术成熟度达到第三层级时，就可以支撑应用能力成熟度的第三和第四层级了，至于第四技术层级是一个技术上更加完备的等级，通过将数据服务产品化为终端用户提供更加高级和便利的服务。