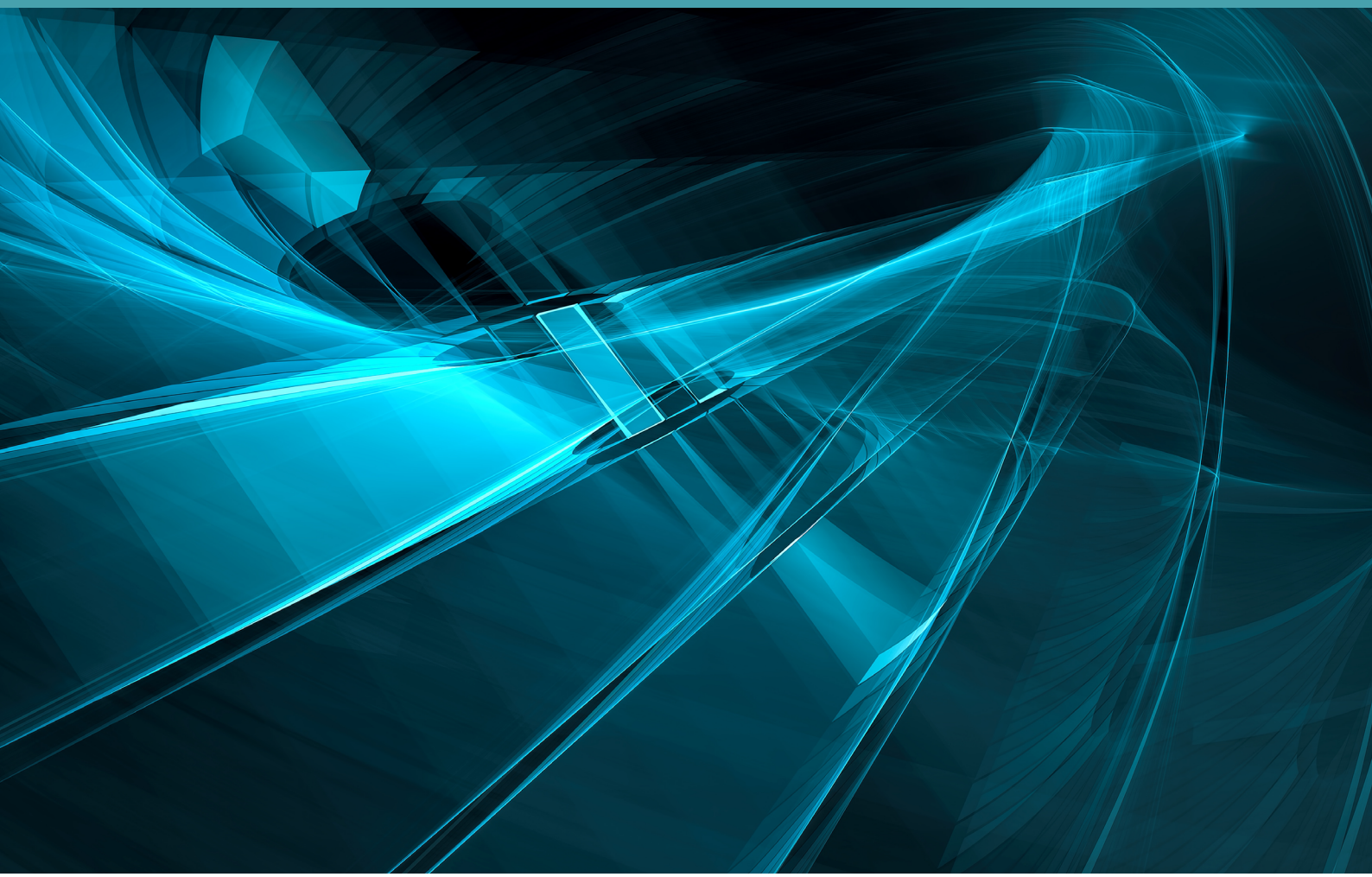
Data and AI



数据工程白皮书

thoughtworks

前言

随着企业数字化转型的不断开展，企业对数据越来越重视、对数据的诉求越来越丰富。本白皮书旨在讨论如何从工程化的角度加速数据到价值的转化过程、为企业带来更多的价值，帮助企业在数字化转型过程中应对来自业务、外部市场、内部数据能力提升等一系列问题。

目前数字化转型对于市场来说并不是一个新鲜事物，从技术视角来看，人工智能与大数据相关技术仍处于创新阶段，各行业正在寻找和探索价值场景与新兴技术融合的平衡点，希望在新兴技术的加持下能够在激烈的竞争中占据有利位置。

近几年企业在数字化以及数据工作上的投入是非常可观的，可是在数据层面上的收益并不尽人意，我们经常听到企业提到：

数据项目投资收益周期长，不确定性大，且没有获取对等的业务回报

数据平台层产生价值的速度跟不上业务需求变化的脚步

数据年年治理，年年治理不好

人员培养困难，培养人才的时间和成本居高不下，无法规模化地支撑业务需求

站在企业的视角，结合 Thoughtworks 近些年服务的客户以及对市场的持续观察，我们发现，超半数的企业认为大数据产业规模将逐步扩大，虽然在部分领域的增速会出现放缓的情况，但是在新兴领域内的大数据产业规模仍将保持可观的增速。另外，大部分企业认为大数据和人工智能领域在未来是值得投资并且能够帮助企业提升效率与客户体验的。

对于企业中需要直接面对或完成数据工作的负责人来说，上述问题需要有一套切实可行的方式方法来确保数据工作能够保质保量的顺利开展、保障企业在数据领域的投入能够有价值产出。这样一套行之有效的方式方法我们称为“数据工程”，而该体系的落地过程称为“数据工程化”。

本白皮书将从实际问题出发，围绕数据工程的定义、实施步骤以及数据价值如何通过数据工程实现等问题展开讨论，并结合当前行业发展态势、数据领域最佳实践、国家数据政策描绘出企业数据工程发展路径，最终对未来数据工程发展趋势进行畅想，进而帮助企业认识自身数据发展现状，制定数据发展规划策略。最后从数据工程化评估的视角讲述如何评估企业目前数据工程化的能力以及如何度量数据工程的优劣，并通过数据工程体系为数据工作者带来启发。

内容简介

本白皮书共分为 4 章，从逻辑上可以分为 3 部分：

- **第 1 部分为第 1 ～ 2 章**

  第 1 章介绍了数据作为新兴生产资料在企业中逐渐占据更重要的位置，第 2 章从企业视角描述了数据在企业中遇到的问题以及数据工程的概念

- **第 2 部分为第 3 章**

  从落地实践的视角详细描述在企业中数据类项目落地过程中的实践，以及如何在落地过程中做好数据类项目

- **第 3 部分为第 4 章**

  基于对数据、数据项目、数据技术未来的发展前景的视角，畅想了未来数字世界中的发展

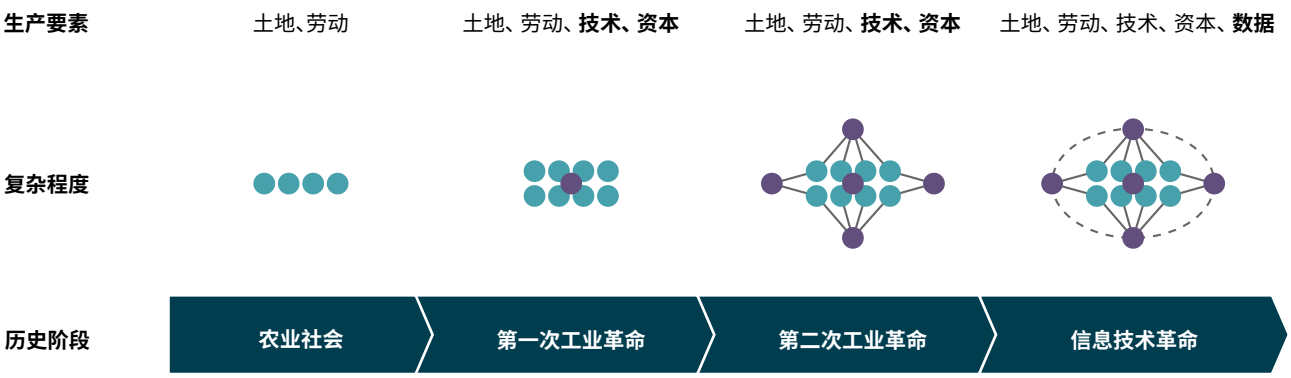
引言

数据是新一代技术革命下的生产要素，掌握了生产要素与生产要素的加工方式就是掌握了数字经济下的价值密码。

新生产要素的崛起——数据

纵观历史，伴随着科学技术的发展以及社会形态的演变，在社会发展的不同阶段，生产要素的数量不断增加，并且每个历史发展阶段，不同生产要素的重要程度也一直在发生变化。在信息技术革命到来之前，社会经济学公认的四大生产要素分别为：土地、劳动、技术和资本，而随着信息技术革命的到来，数据的产生与应用已经渗透到各行各业的生产经营活动之中，数据已经成为继土地、劳动、资本、技术之后的第五大生产要素。数据之重要，已不单单局限于企业内部的认知，更是成为全社会的共识。在 2021 年 11 月工信部发布的《十四五大数据产业发展规划》当中，更是把数据要素的价值转化提升到了国家层面，进一步突出了数据作为国家基础战略性资源的重要地位。

图：生产要素在不同阶段的变化



在当今数字经济时代，一方面企业在经营的过程中时刻都在产生大量数据，这些数据从业务过程中产生，并蕴含着大量知识；另一方面，面对如此重要的生产要素，很多企业无法将其好好利用。而数据又与土地等生产资料不同，其有比较强的时效性，也就是数据对生产的促进作用会随着时间的推移慢慢降低，如果不及时利用将会导致数据价值白白流失，这对企业是一种极大的损失。

而要搞清楚数据价值如何落地，就必然要分析数据的生命周期，包含数据的产生、收集、存储、传输、处理、应用等多个阶段，搞清楚数据从哪里来，到哪里去，怎么使用。数据全流程的不同阶段，需要依赖各种信息系统进行落地，我们将落地过程中涉及到的工程实践统称为数据工程。数据工程的好与坏，直接关系到企业内部数据价值转化效能。接下来我们将系统地介绍数据定义、数据工程的定义、数据工程实施原则。

什么是数据

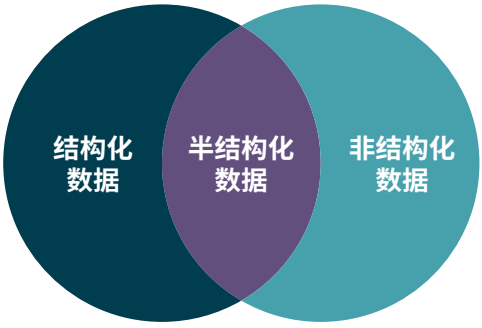
通常，数据是通过观测得到的数字性的特征或信息，是一组关于一个或多个人或对象的定性或定量变量，数据不仅指的是数字，还可以是有意义的文字、字母、符号的组合，也可以是图像、图形、视频和音频等。通常而言，从数据组成形态的视角，我们可以将数据分为结构化数据、非结构化数据、半结构化数据三种。

- **结构化数据**：通常由明确定义的信息组成，这些信息可以通过高度组织化的表格或数据库进行搜索、维护或跟踪。常见的结构化数据如关系型数据库中的客户数据、订单数据、产品数据以及由人工维护的 Excel 表格等。

- **半结构化数据**：是结构化数据的一种特殊形式，它没有固定的结构，因此它不遵循表格数据模型或关系数据库的格式，但是它包含了一些易于分析的结构化元素，例如标记。

- **非结构化数据**：是指没有固定组织原则的未经过滤的信息，如图像、视频、音频文件以及文本信息等。非结构化数据的形式多样，无法用关系数据库储存，且数据量通常较大。

图：不同结构数据的关系



总的来说，结构化数据、非结构化数据、半结构化数据的最主要区别在于是否存在预先定义好的数据模型。结构化数据能够用统一的某种结构加以表示，离开了这种结构，数据就没有意义；非结构化数据没有概念数据模型形式的限制，可以自由表达；半结构化数据介于上述两者之间。

数字化转型浪潮下的企业数据

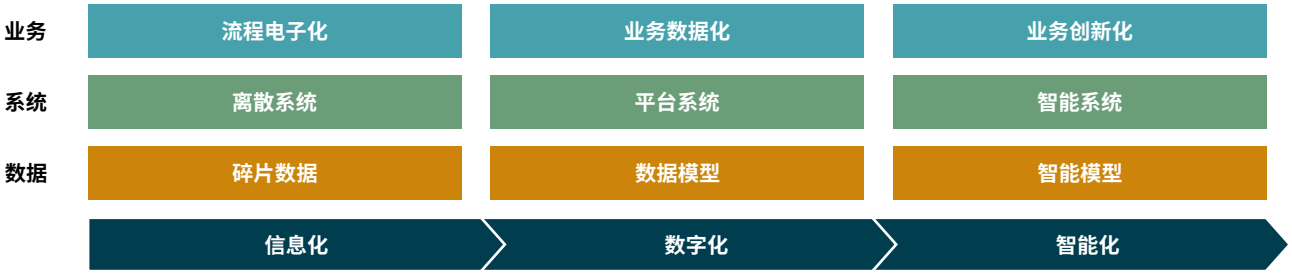
在了解了数据具体含义的情况下，企业想要更好的管理数据、利用数据，就必须了解数据在现代企业中的产生源头、组织形态等。现代企业数据的产生离不开企业的数字化转型，企业数字化转型程度高低则直接影响了数据的利用效率，在分析了众多企业数字化转型的案例之后，我们认为企业数字化转型一般分为三个阶段：

- **信息化：**信息化为企业数字化转型的初级阶段，此阶段侧重于将企业生产制造过程、物料转移、事务处理、资金流动、客户交互等流程进行电子化，其整个思维导向以流程管理为主，以无纸化办公为目标，旨在提升企业流程管理效率，这一阶段的企业主要呈现为系统离散化，数据碎片化的特点。

- **数字化：**在企业信息化达到一定程度之后，由于业务的快速发展，原有流程和系统已经不能满足企业的管理诉求，企业逐渐由流程管理转向业务管理，企业对其业务进行细粒度的拆分、分析与优化，便于对制造流程、业务流程、用户旅程等进行管理、分析与改善，这一阶段为企业数字化转型的中级阶段，主要强调数字对商业的重塑，转型过程中通常伴随着组织结构的调整，赋能企业商业模式不断创新和突破。处于这一阶段的企业信息化主要呈现系统平台化、数据集约化与模型化的特点。

- **智能化：**在企业拥有大量数据的背景下，伴随着人工智能领域技术的快速发展，原本只在学术界活跃的人工智能算法与模型能够快速在商业领域落地，智能算法与模型极大提高了企业从数据中提取业务知识的效率，企业各种系统与应用变得越来越智能，系统在算法与模型的帮助下可以自学习知识、再创造知识。智能化由于天然的高效，成为了企业数字化转型的必然趋势，此时系统构建的思维导向为业务创新，旨在利用人工智能算法与模型解放生产力、寻找新商机。此阶段企业主要呈现为系统自动化、数据模型化与智能化的特点。

图：企业数字化转型三个阶段



企业数据流转链路分析

数据只有依托于信息系统，才能在企业内部流转起来。数据在企业内部会经过一系列的处理才能最终产生价值，一般我们会简化为以下几步：

- **数据产生：**一般而言，企业数据由生产活动以及服务客户的过程产生，不同行业的数据产生特点也不相同。如果是生产型企业，数据主要由传统信息系统如 ERP、CRM 等系统产生；如果是服务型企业，则数据主要在不同类型的在线系统产生，例如电商系统、推荐系统等。除此之外，由于物联网的发展，还有一部分数据由传感器产生。此时数据还相对原始，其形态可能有结构化数据、半结构化数据、非结构化数据。

- **数据收集：**数据收集通常是指将业务数据从业务系统或者外部系统接入的过程。数据收集阶段需要满足三大原则才能保证后续步骤的有序进行：首先是无侵入性原则，通常数据接入需要直接对接业务系统，而对业务系统而言最重要的是稳定性，即数据收集过程不能对业务系统造成负担。其次是无修改原则，数据收集是数据工程所有流程的起点，在数据收集过程中数据一定要与源系统保持一致，避免不必要的处理导致数据所蕴含的信息缺失。最后是可追溯原则，收集来的数据可以进行冷热备份，但不进行任何删除操作，便于审计、回溯等。

- **数据预处理：**收集的数据格式存在多样性并且掺杂着有效或无效的数据，导致这些数据无法直接进行利用，必须要进行相关预处理才能进入下一阶段。这一步骤会提升数据的信息有效密度，并且会对数据进行转换与处理便于后续计算，一般而言数据清洗、数据标注、编码等均属于这一步骤。

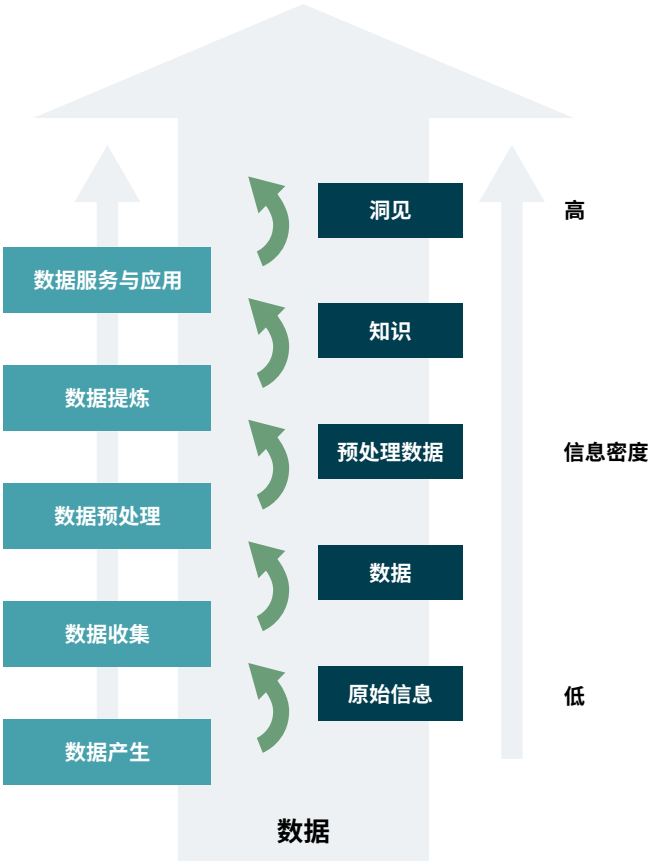
- **数据提炼：**此阶段为数据价值转化的主要步骤，从数据中提取信息、凝练知识就发生在这一步。对于一般的数据仓库而言，数据模型建立、ETL 计算，以及业务标签构建，都在这一步完成。而对于机器学习类的平台，智能模型的训练也可以归到数据提炼中去。

- **数据服务与应用：**经过逻辑计算完成后的数据，蕴含了大量的信息，是指导决策的重要依据。通过提供在线数据服务或者应用的方式，使得数据价值能够自动、高效落地。我们常见的数据 API、BI 报表、AI 模型的在线应用都属于这一范畴。

- **数据治理：**数据治理严格来说并不能算作数据生命周期中的某一环，数据治理是贯穿整个数据生命周期的。为保证企业内各个业务领域数据工作的有序开展，就必须对数据进行统一的规划，包括数据资产、数据标准、数据质量、元数据、数据安全与隐私等，我们将这些工作统一划归到数据治理的范畴之中。

从数据产生到数据价值落地的过程中，数据的信息密度越来越高，其中蕴含的知识也越来越丰富。虽然并不是所有的企业在数据工程落地过程中都需要对数据全生命周期进行分析与管理，但是如果不去分析数据的全生命周期，很容易导致“一叶障目不见泰山”，那么就必然会出现企业对数据认知不足、规划不清晰的情况。通过分析企业数据全流程，企业可以识别薄弱环节，抓住重点环节，因地制宜的制定数据工程落地规划，所以说数据全流程分析，是每个企业在进行数据工程落地之前的“必修课”。

图：企业数据流转链路



数据工程概述

随着数据重要性的不断提升、数据在企业内的流转越来越常见。数据工程则是帮助企业高效地挖掘数据价值，持续地赋能业务增长，加速数据到资产的升华过程的最佳实践。

数据在企业流转的问题

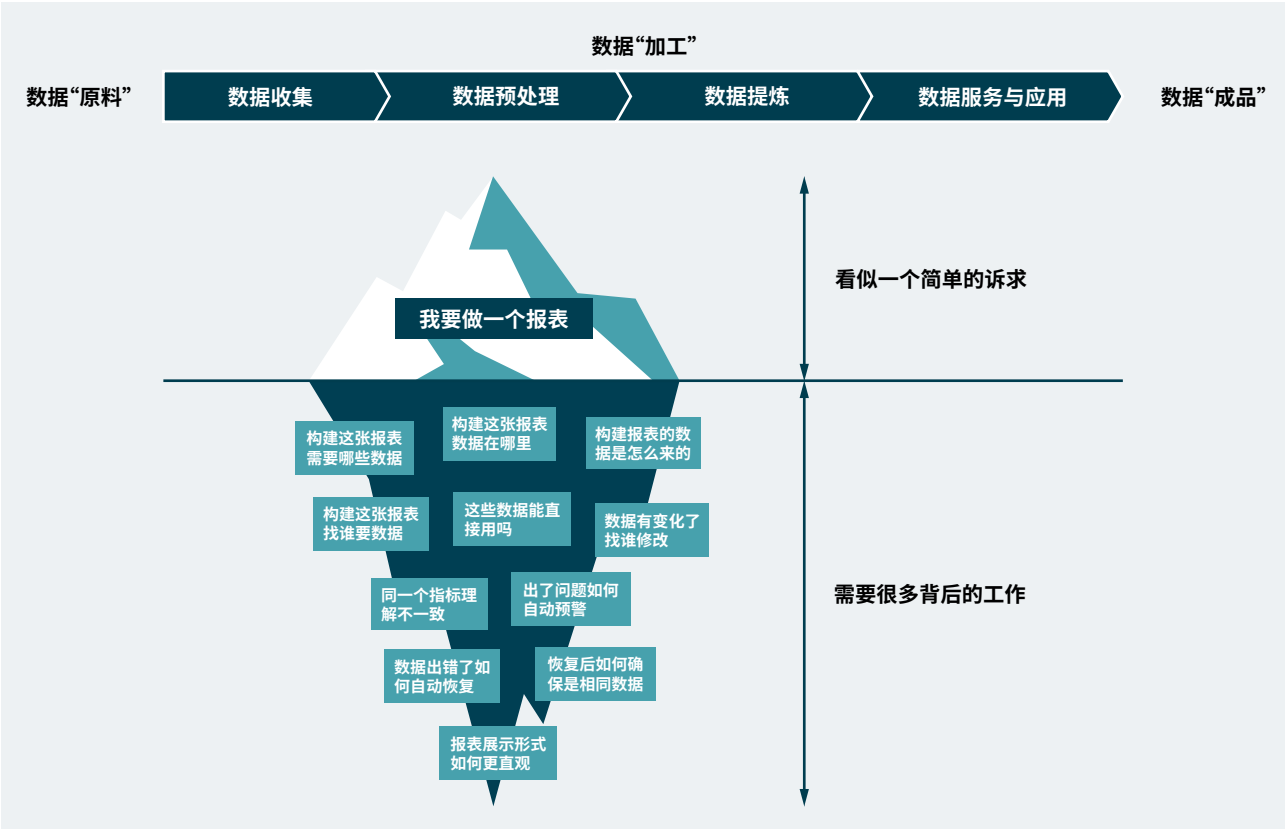
数据在企业内部流转会经历多个阶段，而每个阶段之间还存在着各种各样的问题。数据是用来产生价值、为企业提供便利的，因此企业的发展阶段、企业对于数据使用或产生价值的诉求也有不同，想要解释清楚数据在企业流转的问题，就不能忽略企业自身的诉求和特征。接下来我们将展开来描述这些阶段以及区别。

首先，我们先来看企业通过数据想干什么，企业可以分成以下四个阶段：

- 通过数据描述企业正在发生什么，清楚地了解企业的“数据形态”。
- 通过数据回答企业为什么正在发生这些变化，探明企业遇到的问题、发展的变化都是由什么引起的。
- 通过数据帮助企业在遇到问题时进行示警，明确下一步行动的方向在哪里。
- 通过数据帮助企业应对数据展示出的业务状态，动态调整投入以确保得到预期的产出。

从上述描述中不难看出，在不同的阶段，数据都可以为企业带来价值，这些价值产生的过程就是数据在企业内部流转的过程。为了方便理解，我们以做报表为例看看数据的流转都经历了哪些过程，相信不少数据从业者都经历过类似“手工 Excel 维护表格只需要 2 小时，为什么要花 3 天时间做报表”的灵魂拷问，这里的 2 小时也好 3 天也罢要从实际的诉求出发，仍有企业仅需手工维护的方式就能满足诉求，因此我们也并不推荐为了做报表而做报表，这里要讨论的是对数据展现过程自动化、低廉运维成本、数据可信、报表直观可用有诉求的企业。

图：报表实现过程中的数据流转



如上图所示，数据从“原料”到“成品”并非是简单的将数据接进来、展示出去，而是将数据自动化地从系统中获取到、按照业务逻辑对数据进行补全和纠错、通过统一的各部门都认可的计算逻辑来进行计算、用更友好和直观的方式将数据展现出来。

数据企业流转的过程中，收集、处理、计算、使用这几个核心的步骤仅会因为企业对于数据不同的诉求而导致这四个步骤实际处理起来的复杂程度有所区别。因为企业的实际情况不同而导致这四个步骤实际处理起来有所倾斜，但总的来说并不会因为这些区别导致其中某个步骤被舍弃。因此，企业收集、处理、计算、使用的过程有快慢之分，这快慢之分的核心就是企业在数据工程实践好坏的区别。

什么是数据工程

正如前面提到的，数据工程能够加速数据接入、处理、计算、使用的全流程，但是对数据工程到底是什么缺少一个清晰的描述。

要解释数据工程是什么，就需要从软件工程说起。从软件开发出现到软件开发逐步规模化的过程中，IT 从业者们一点点积累下关于需求、设计、实现、测试、运维等方面的工作最佳实践，因此我们不难看出软件工程并不仅仅是软件开发，而是一套体系。这套体系不仅仅适用于应用开发，也同样适用于数据开发，因此数据工程是软件工程的一部分。这里为了方便大家理解，我们将软件工程从产出物类型的角度划分为数据类和应用类。对于数据类产出物的需求到运维的规模化全过程体系就是本文中讨论的数据工程。

数据工程包含了需求、设计、构建、测试、维护演进等阶段，涵盖了项目管理、开发过程管理、工程工具与方法、构建管理、质量管理，是一套为了应对规模化生产和使用数据、为业务提供数据支撑，最终产生价值的体系。同时定义了在落地实施过程中如何确保需求准确性、设计灵活性、开发便捷性、维护低成本性、架构可修改性等保障性能、质量的原则。

总的来说，正是因为数据有着不同的种类、不同种类数据处理有着不同的特征，让我们对上述定义再换一个角度来审视：

- 数据工程是一套体系
- 数据工程是用来加速数据到价值过程的规模化最佳实践
- 数据工程是软件工程的一部分
- 数据工程不是传统软件工程在数据领域的简单重现

数据工程价值

数据工程并不是单一的大数据系统或平台的落地，因此数据工程的价值并不能仅从普通的信息系统的角度来看。数据工程的好与坏，往往与企业的组织架构、团队协作、实施能力等息息相关。而针对企业所处数字化转型的不同阶段、所处行业业务特点以及企业本身组织架构，数据工程价值凸显的点也往往不尽相同。我们自顶向下详细分析了优秀的数据工程能够在不同层级给企业带来的价值，方便企业找到自身在数据利用上的主要矛盾。

在企业层面，数据工程的实现从业务出发，在企业层面打造高响应力且更加智慧的业务，加速从数据到价值的服务产生过程。数据工程化的实现，能将分散在企业内部各业务系统中的信息流数据进行融合、打通，对内实现共享的数据入口进行统一化、标准化。同时，标准化的入口支持企业外部系统或数据的快速接入。通过收集、汇总、清理、结构化、存储，达到数据治理的效果，并实现数据溯源。它能将企业发展不同阶段的分散数据进行汇聚，将数据价值构建成各种服务支撑业务，对外能够更好地服务企业客户，实现真正的“以客户为中心”。最终数据工程可以挖掘数据的价值，帮助企业创新业务、提高效率，将数据从成本变成资产。

在团队层面，数据工程可以实现减少内耗，提升效率，解决数据开发与数据产生价值的协作问题。可以在满足企业各部门自身需求的同时，统一数据标准、解决数据孤岛问题，降低各业务的联动成本，提供组织内部的协作，支撑业务快速响应。可以更科学地构建整体架构，实现基于中台的数据统一，真正为业务创新和服务带来价值。

在人员层面，良好的数据工程实践可以降低人员成本，解决很多企业的开发人员、技术人员没有数据能力的问题。通过集中地对跨部门数据的采集、融合、治理、组织管理、智能分析，可以大大缩减人员规模，降低人力成本。一致化的工程实践可以提升开发质量阈值，降低开发人员的理解难度，解放运维工作，让开发人员更专注于业务价值。

数据工程的价值体现需要有价值体系来度量，而这套价值度量体系则依赖于数据工程在落地实践方面的具体内容，因此接下来将详细展开介绍数据工程落地实践。

图：数据工程的价值体现

数据工程落地与能力建设

数据工程在企业内部带来不同层面的价值，做好数据工程、让数据工程能够在企业内落地，形成匹配企业特征的能力体系是关键。

图：数据工程价值观及原则

数据工程价值观		
根据数据产生的价值作为交付结果	胜过	根据数据接入、数据处理、指标计算数量等作为交付物结果
全功能团队协作的端到端开发	胜过	按数据处理流程的分段开发
按业务域划分的面向未来的设计	胜过	按技术堆叠的限于当前的设计
团队的知识积累和传承	胜过	简单的文档交接

数据工程的7条原则
功能设计与开发要从价值交付考量
合理的架构设计不仅指解决现有问题，还能够在一定程度解决未来问题
我们倡导通过统一的工作标准和流程提升团队协作效率
工具是知识沉淀的具体表现，有效的工具能够提升规模化开发效率
欣然面对需求变化，及时调整交付策略
数据治理需要渗透到整个数据工程落地过程当中
人是数据工程落地的核心，要注重人员培养、知识传承

数据工程落地

在面对业务协同性不够、业务决策路径不清晰、组织架构可能导致的部门墙等诸多问题上，我们期望将企业多业态、多链路中所涉及的不同业务数据汇聚、打通全产业链、构建业务生态，打造以数据为中心的价值创新产品，通过数据去产生新洞见、发现新业务、打造新产品、验证新想法，从而驱动业务的快速迭代。

对于企业来说，我们推荐三步走战略：数据愿景对齐、数据工程落地实施、数据持续运营。三步自顶向下，先确定总体目标，再进行目标拆解，由目标制定具体措施，再到具体工程实践，最后以持续运营手段，完成数据从业务中来，再到业务中去的完整价值闭环。数据愿景对齐作用主要是明确企业数据愿景，保证后续步骤不偏离企业本身的价值实现，主要包括业务场景价值的的探索识别、优先级评估、数据架构设计、技术架构设计等。落地实施主要包括数据平台的建设落地，如数据的采集、清洗、存储、计算、测试等。持续运营则是为了保证在数据平台建成后能够及时响应变化并做出调整，源源不断从数据抽取价值来反哺业务，最终实现愿景。

图：数据工程落地三步走战略

愿景对齐

回顾 Thoughtworks 在对上百家企业进行数字化转型的咨询与交付中，我们发现由于所处行业特色、企业组织架构、数字化转型成熟度以及企业规模等不同，导致企业对于数字化转型的愿景并不相同，有的企业数据愿景注重数据应该被如何共享、数据应该如何协作使用；有的企业更关注数据服务如何更快、更好、更智能的服务于业务系统；而有的企业则更关心数据质量如何保证、数据标准是否统一、数据管理如何更简单高效等问题。因此在前期数据战略中拉齐愿景就显得尤其重要，不然会舍本逐末，过分追逐于解决某些具体问题，忽视了企业宏观目标的把控。

在进行数字化转型过程中，前期的战略规划准备不足或设计不合理，都会导致后续落地无法正常进行。在过去的数据战略中，通常解决的是企业数据管理问题，目标是服务于 IT 战略，让数据管理更规范，服务于企业管理，而不关心客户。因此其核心目标就是管理好数据，如何进行数据的清洗以提升数据质量，如何进行数据的管理认证以确保数据的权威性和有效性，如何对数据进行权限管理控制以解决什么样的数据可以被什么角色什么部门进行使用等，所以过去更多的是从企业内部视角来做数据战略。同时，传统的数据战略通常是以管理大而全的数据资产出发，围绕企业内部组织、流程规范、规章制度，以数据现状为基础进行战略规划，但往往会面临以下问题：

- **缺乏科学方法论：**规划制定是个复杂的过程，需要团队有全局观察能力以及详细方案制定能力，而往往企业内部由于数据团队与业务团队相对割裂，懂数据的人不懂业务，懂业务的人不懂数据。在短时间内没有相关赋能者的时候，如何应用科学方法论就显得及其重要了。

- **规划不合理：**在数据工程落地规划过程中，业务价值实现是最终目的，技术方案落地只是手段，规划一定要分清主次。由于落地过程规划不合理，导致企业了花费大量人力物力，可是投资回报却差强人意，整个价值链不完整，导致半途而废。

- **缺乏验证手段：**全局的、端到端的覆盖整个企业的规划是个浩大的系统工程，不仅周期长，而且短期内很难看到效果，许多企业在建设过程中缺乏耐心导致最终放弃，或者虽然有了阶段性成果，但由于缺乏统一衡量标准，价值无法验证，目标也只能是空中楼阁。

通过上述问题我们不难发现，企业需要科学方法论负责统筹规划，落地规划拆分成实施步骤，验证手段则用来评判结果，所以应对上述问题的思路就是愿景对齐。

图：愿景对齐的四个步骤

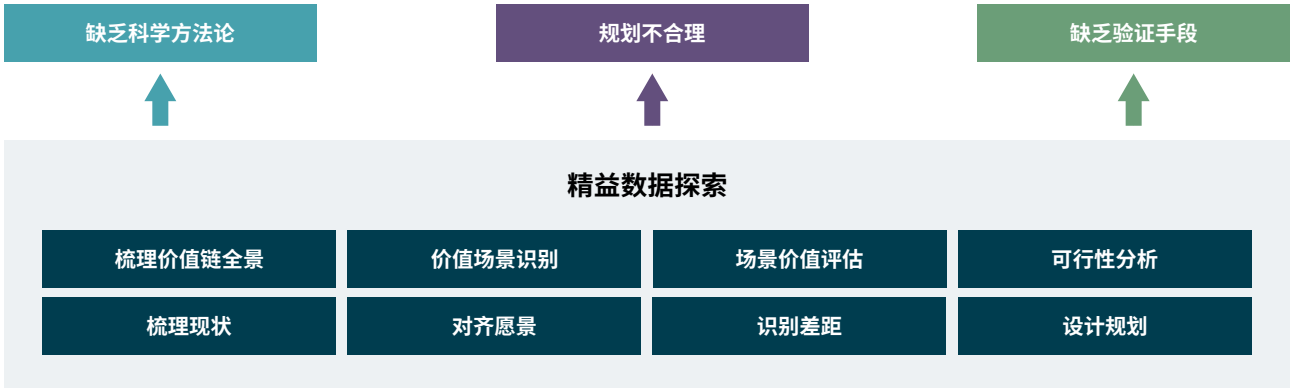
愿景统一，业务价值的场景探索。愿景对齐的第一步要素就是价值场景的探索与识别，通过定义、统一业务价值度量框架来识别业务价值场景。这一步面临最大的挑战是业务和技术的鸿沟，由于业务人员在业务价值场景探索阶段不清楚哪些技术会更好的解决问题或产生哪些价值场景，从而会局限在自己的认知中，很难探索出业务与技术融合的、可落地的高价值场景，而技术人员对业务不了解也会导致探索的价值场景不被业务认可。那么如何更好的将业务和技术结合以探索出更多更有价值的业务场景就是核心关键。具体来讲，通常先梳理当前数据现状，如当前数据模型有哪些、数据质量如何，业务价值场景是否有数据支撑；再梳理数据要给谁使用，通过分析不同数据角色的数据用户旅程，梳理出数据用例、数据价值流。从而在业务场景中将数据的消费方，生产方及数据全生命周期的蓝图构建出来，再引入技术人员的技术手段，对上述梳理数据的排列组合来进行创新性的头脑风暴，即围绕业务愿景对物理世界的业务构建出数据全景，通过业务模型之间的组合发散业务场景，从而产生创新的业务价值场景。在场景探索结束后，需要有对应的业务价值评估体系来对场景进行业务评估，基于解决的痛点和产生的价值权重来进行价值评估。通常，探索出的业务价值场景需要包含场景的背景、价值点、所涉及的用户、需要什么样的能力、用户旅程、所涉及的实体、风险等信息。

价值、成本、可行性的优先级排序评估。在业务价值场景探索完成后，接下来就需要评估数据质量、技术可行性、业务痛点，辅助战略目标来产生价值优先级排序策略。因此在对优先级的评估阶段，是从可落地的视角出发，以防止前期探索得出的业务场景变成空中楼阁。在业务价值场景探索和优先级排序完成后，需要对业务、系统、痛点、数据成熟度等一系列的现状产出，从而为后续的架构设计提供输入。

合理的架构设计。在架构设计阶段，要考虑如何将数据产生的价值规模化，从数据的接入、处理、使用等数据全生命周期流程中所涉及的业务扩展性、时效性、安全性、可复用性、便捷性等原则，从而进行架构设计。通常，架构设计包括数据架构、技术架构、安全架构、数据治理规划、数据运营策略等。

整体规划的方案制定。当上述价值场景、优先级、架构蓝图都梳理清晰后，接下来就是定制项目计划、快速启动建设，分阶段的定制路线。最后，需要有项目规划设计的成熟度评估。

图：愿景对齐挑战的应对策略



落地实施

愿景对齐固然重要，但是如果做不好数据工程落地，那么愿景与规划都是空中楼阁。在具体的工程实施中，每个系统又都是一个自顶向下设计，自下向上实施的过程。其落地过程就如同孕育新生命一般，其中数据梳理规划蓝图，数据架构设计规划骨架，数据模型设计构成器官，数据接入则赋予信息感知能力，数据处理构成中枢大脑，测试、安全部分负责为新生儿提供保护，每个步骤相互依赖，缺一不可。所有这些有机组合才能完成数据全链路管理，智能高效地实现数据价值转化闭环。

本章节将通过数据梳理、数据架构设计、数据接入、数据处理、数据测试、数据安全和能力复用与保障七个步骤来描述在数据工程落地过程中所要遵循的原则规范。

数据梳理

在实施过程中面对的是不同种类、不同特征的数据，某些场景而言数据梳理可以是单独的项目或者是其他项目的前置步骤。数据梳理就是要全域分析数据粒度，规划数据层次以及统一数据口径。这么做的目的是整理清楚数据所代表的业务含义、去除跨部门和跨场景在理解上的不一致、寻找使用数据和计算的统一口径、找到能够维护数据的管理者，最终构建在企业内部能够描述数据流转过程、数据变化过程的全景。这么做的好处是让数据使用者能够对数据的变化有全面的认识，对于后续数据项目开展提供扎实的基础。

如前文所述，数据的背后是信息、是业务知识，因此我们想要理清楚有哪些数据，就需要先对业务流程进行梳理，根据项目类型的不同需要梳理的业务流程范围也会有所不同，比如：围绕整个公司视角的梳理、围绕某个场景的梳理，但无论是哪种范围，都需要把业务流程梳理出来。业务流程的梳理仅仅是第一步，业务流程梳理的目的是在于产出基于业务流程关键节点有哪些数据，通常来讲我们需要精确到字段级。对于数据工程而言数据梳理可以从以下视角来审视：

图：数据梳理的三大目标

数据分级分类	统一口径	确定数据owner
主题域大类	业务含义统一	数据流转图
主题域	指标口径统一	实体属性所有者
业务实体	业务边界统一	
实体属性	类型统一	

- **数据分级分类：**面对企业多业态、多链路复杂流程的场景下，会涉及不同角色不同部门的不同级别和类别的数据，因此在前期我们需要对齐数据的分级分类。数据梳理的核心其实是领域模型、实体模型和业务流程的梳理，需要从组织架构、业务流程等进行主题域的分组划分以及确定所涉及的实体和实体属性的信息。分级分类一方面可以更好的理解业务和数据，从而更清晰的得到数据全景图，为后续的数据处理和使用做准备，另一方面可以了解其数据分布，在运营阶段更好的进行数据管理。此外，基于数据的分级分类，可以更清晰的划分数据边界，帮助业务更好的梳理和优化业务流程。同时，也需要基于安全的视角对数据进行分级分类，从公开数据、内部数据、机密数据等级别进行划分，从而决定后续的数据共享策略。

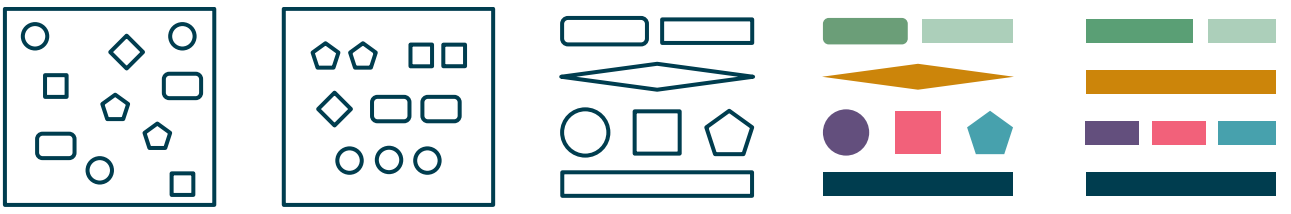
- **统一口径：**在上述梳理完数据的分级分类后，应该已经对整个业务流程所涉及的实体有了清晰的认知，那么口径的统一是在统一什么？这里提到的主要是实体的口径统一和实体内指标的口径统一。对于实体的口径，在业务系统的设计开发阶段，通常都是围绕业务流程进行，也就意味着并不会过多考虑同一个实体跨业务系统的定义，导致同一实体在不同业务系统的业务定义、业务边界等不相同，但是口语间的业务传递描述又是相同的实体，即相同现实世界中的实体在数据视角下的业务定义和边界可能不同。实体的边界划分通常是基

于业务决定。对于指标的口径，通常在使用数据进行分析或数据挖掘时，指标信息的业务逻辑定义就尤为关键，在业务复杂的场景下，指标信息的定义从大分组上定义相似，但是又有细微的逻辑差别，如利润的定义在不同的企业中就有多种细粒度的划分，在数据的使用阶段，就需要更加清晰和统一其指标信息。

- **约定数据 Owner：**在业务流程中，不同的部门和系统会使用已有的数据，并可能会对已有的数据在某个业务流程的节点上进行修改，同时也可能基于现有数据产生新的数据。那么面对多版本、多边界的实体数据，如何保证使用数据的部门和系统所使用的数据就是所期望的数据呢？因此我们需要进行数据的 owner 梳理。这里与其说是梳理数据 owner，倒不如说是梳理业务流程中不同实体的生命周期变化的关键负责人是谁，如在什么时间什么业务背景下谁对什么实体的什么属性做了什么修改，为什么要这么做等。当然这里所讲的数据并非一个实体，而是会细粒度到实体的某个属性，甚至是某个属性的某个值，如订单状态的值。同样，到底是粗粒度的实体还是细粒度的属性值定义边界，依然是由业务决定，即是基于业务流程中的核心节点来决定。通常来讲数据 owner 与数据在映射管理关系是一个一对多的过程，即一个数据 owner 会负责至少一个数据或者是一类数据。企业根据数据 owner 所处的部门、负责的业务域、所对接的业务部门、所处的权限级别，可以将分级分类后的数据域数据 owner 进行映射，形成企业自己的数据管理体系。数据 owner 需要定义数据的业务含义、业务边界、数据标准和数据的使用权限等。

- **构建数据标准管理流程：**我们知道了要找谁来修改数据，可是如果数据被修改错误、或者是修改的不符合业务场景和标准，可能会引发一系列新的问题。我们约定数据管理者的初衷是能够让数据得到正确的修改，而不是引发新的问题。因此我们需要的是让数据管理者根据技术对数据的要求、业务对数据的要求对数据进行修改，所以构建的数据标准管理体系要包括数据标准、数据安全权重。到目前为止，我们有了管理数据的人、管理数据的方式，我们就拥有了可用的数据，无论是将数据提供给其他系统还是为即将开展的项目提供数据基础就已经具备一定的基础了。从数据使用的视角来看这些数据可以通过集中管理的方式来提供出去。

图：数据梳理的过程



数据架构设计

数据架构是一个比较泛指的概念，当前权威组织对于数据架构内容的定义也有所差异，如《数据治理：工业企业数字化转型之道》所提到的数据架构包括数据主题域、数据关系、数据分布、数据模型等，《DAMA 数据管理知识体系指南》提到的数据架构指的是数据模型和数据流设计，而这里我们说的数据架构设计主要从数据存储模式、时效性和分布模式三个架构设计和数据建模视角来描述。

图：数据架构设计

存储模式	时效性	分布模式
数据仓库	流处理	集中式架构
数据湖	批处理	分布式架构
湖仓一体	流批一体	

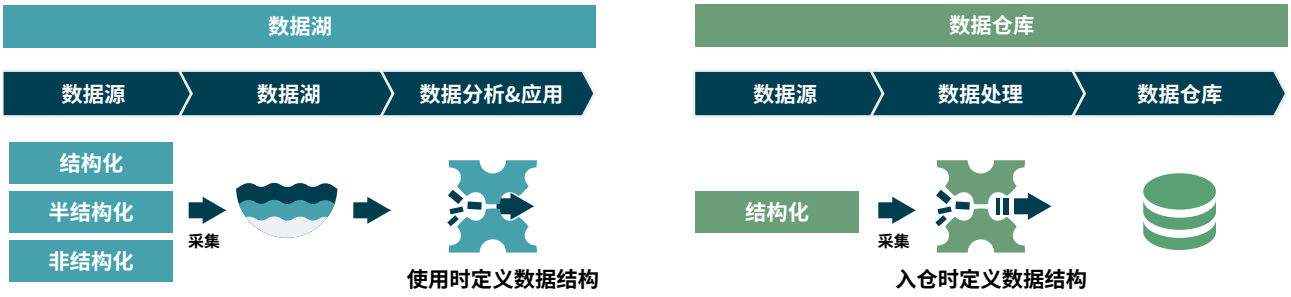
数据的存储模式划分主要可以分为数据仓库和数据湖两种。数据湖是一个集中存储区，用于存储、处理和保护大量结构化、半结构化和非结构化数据，可以基于事先定义好的 schema 来对数据湖中的数据操作，可以总结数据湖的特点如下：

- 集中式存储库。
- 保持原始数据格式而无需对数据进行处理。
- 支持丰富的计算模型。

而数据仓库是用于分析结构化和非结构化的数据，通常数仓的数据已经定义好其 schema，总结数据仓库的特点如下：

- 内置的存储系统，不会暴露原始的数据源文件。
- 通常需要通过 ETL 或者 ELT 对数据进行清洗和加工。
- 更加侧重数据建模和数据管理，供商业智能决策。

图：数据湖、数据仓库简介



数据湖设计是通过开放底层文件存储，给入湖的数据结构带来最大的灵活性，结合上层的引擎，可以根据不同的场景来随意读写数据湖中的数据，并进行相关分析。缺点也很明显，缺少模型导致对业务的处理的维护成本以及随着数据规模的不断增大而增大，最终可能变成数据沼泽。而数据仓库设计更关注大规模业务数据下的数据使用效率和数据管理，通过数据模型来保证对业务的理解以及通过模型复用来保证数据的使用效率。缺点就是在前期数据仓库搭建阶段的数据建模成本较高，周期较长。

因此，在对于这两种技术架构的设计，需要根据企业的不同需求来选择。对于业务灵活多变的场景，数据从生产到消费需要一个探索性的阶段才能稳定下来，那么此时灵活性就更加重要，数据湖架构会更加适合。而对于业务成熟稳定的企业，则更需要对于数据仓库的架构，来帮助企业沉淀数据的流转、数据处理流程和数据模型等，以支撑不同数据消费者对数据的高效使用。

数据的时效性划分可以分为实时处理和离线处理两种方式。通常，实时处理是以流处理或微批的方式体现，而离线处理通常是批处理的方式。

在批处理模式下，我们通常会周期性的对一段时间的数据进行采集和处理，批处理的数据集通常包括有界、大量、时效性低的特点。而流处理通常是事件驱动，因此其处理的对象并非数据集，而是单条数据。因此他们在数据处理时通常有以下不同：

- **时效性。**批处理对于时效性要求不高，通常需求是查看历史某时间段的数据分析等，而流处理的需求则是要求数据发生变更时就需要进行相应的数据处理，以获取最新的数据结果。
- **数据量。**批处理在进行数据处理时，其吞吐量通常比较大。而流处理则是单条数据处理。
- **准确性。**在面对分布式大批量的数据接入处理环境下，为了保证需求的时效性，通常流处理会牺牲部分数据质量要求来满足其数据的实时性。

传统的数据处理通常是批处理模式，但随着业务和需求的发展，会导致某些企业也会逐渐包含流处理架构，这就导致数据的处理会同时包含流批两种模式，因此在处理阶段，我们通常需要遵循以下要求：

- 流处理和批处理的数据处理逻辑尽量使用相同的业务处理逻辑
- 对于流处理，需要考虑数据准确性和时效性的平衡
- 流批处理都需要确保数据的语义一致
- 无论是流处理还是批处理，都需要保证数据端到端的一致性

数据分布模式即数据模型在前期的顶层设计，通常有两种设计方式：

- **面向企业的整体数据设计**，即集中式架构。
- **面向领域的敏捷数据设计**，即分布式架构。

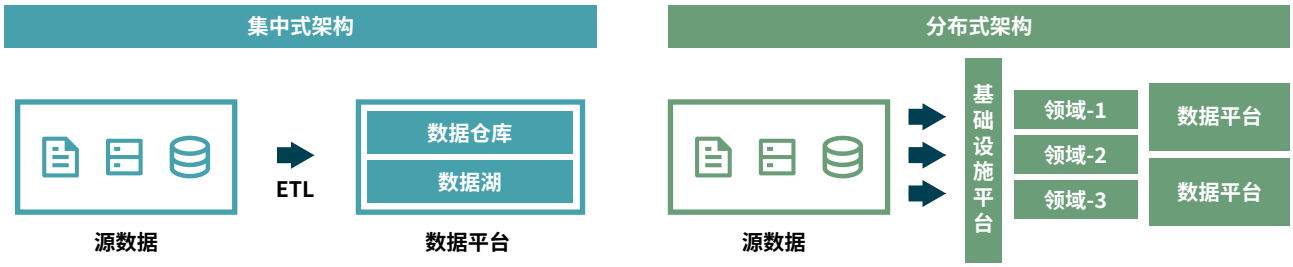
集中式架构是以企业视角进行数据建模，包含了企业内不同领域的数据，而分布式架构的核心则是面向领域的数据建模。以下是这两种架构的区别：

- **数据管理模式不同：**分布式采用分而治之的思想，围绕领域划分，将数据的所有权交给了领域团队，遵从“数据在接近其来源的地方进行管理”原则，可以更好的进行数据治理。此时，领域团队应当同时具备业务和技术能力，能够把领域专业知识与创造商业价值所需的技术能力结合在一起。集中式架构则是需要对不同领域的业务进行系统的学习了解，并结合业务架构来进行系统的建模，因此集中式架构依然是由数据团队来管理数据，业务团队仅仅作为业务的输入方来参与数据工程，此时数据团队应当具备一定的业务领域知识。

- **需求响应灵活度不同：**集中式提供具有内置计算能力的结构化数据存储，满足企业内的所有需求。然而，企业规模越大，这就越不现实。除了最简单领域外，所有领域都需要多个限界上下文以及相应的数据模型，同时，集中式架构下，基础设施资源所有业务共享，进行集中式的管理和维护，无法基于业务需求灵活进行资源调整。而分布式架构中，面向领域的数据会作为平台的最小业务单元，每个数据都具有独立的灵活技术栈选择、可发现、可寻址、自解释、合规、安全、可管理、可扩展、以及相互运营性，从而保证企业在更复杂多变的业务场景中获得最大灵活度且可扩展的数据能力。

- **需求边界不同：**分布式是按照业务领域或者功能来划分进行数据建模，这就导致分布式架构无法直接满足跨领域需求；而集中式是按照企业层面进行数据建模，因此集中式关注点是企业范围的需求识别、规范，因此可以保证数据的多样性，可以实现跨领域的需求场景

图：集中式架构和分布式架构的区别



上述提到的数据湖和数据仓库、流处理和批处理以及分布式架构和集中式架构，这些都是需要基于场景以及需求进行选择，甚至可以在某些场景下混合使用，如我们所熟知的湖仓一体、流批一体等，所有的架构都有其适合的场景和所需要的成本。简言之，没有最优的架构，只有最合适的架构。无论架构如何选择，都不会影响本身的数据建模，那么目前常用的建模方式有三范式建模、维度建模和 data vault，这里我们推荐采用维度建模。

为什么使用维度建模。维度建模重点解决在保证数据质量的前提下，如何更快速的完成分析需求，同时又要保证大规模数据下的复杂查询的响应性能。因此，在数据的明细层，我们需要确保维度表能够包含实体属性的缓慢变化情况，而事实表作为操作型事件的数据表现，需要最真实的反映现实世界的行为，也就意味着，事实表的模型中需要包含更多的业务信息，如包括不同渐变维度数据下的行为表现。这样做的目的是可以通过模型更加直观的发现数据质量问题，保证数据的一致性、准确性和完整性，同时，模型可以快速响应更多的需求场景。但缺点也很明显，在面对需求进行数据分析时，由于需要多表关联，导致成本会比较高。因此，在需求侧我们通常又会将上述明细层的数据进行进一步处理，得到我们可以满足当前需求的模型，从而解决上述分析成本高的问题。我们可以发现，维度建模在不同层所解决的问题不同，最终都是为了保证数据质量、快速响应需求和性能。

数据建模阶段，我们通常会分为概念建模、逻辑建模和物理建模，这三个阶段的侧重点也不相同。概念建模的主要目标是对各概念实体进行归纳和总结的过程，是比较粗粒度地进行业务描述，其主要是定义主题域和实体、实体间的关联关系。逻辑建模的目标是细粒度完整的描述业务场景，为确保其可以最真实的反映现实世界的行

为，需要确保其数据源、数据 owner、数据粒度及其属性的边界等。物理建模目标是基于逻辑建模对业务的认知，进行模型的存储设计，主要考虑技术选型、需求场景、计算存储成本和响应诉求等。基于建模的三个阶段和上述的不同分层，我们可以总结得出建模的几大原则：

- **模型分层：基于不同的设计目标进行分层。**贴源层主要目标是为了记录最真实的源数据，在法律法规允许的前提下，尽可能的保留每个版本的数据，以方便后续运维。明细层结合维度建模主要是为了保证数据可以反映最真实的业务场景，保证数据质量和多变的需求快速响应能力。而服务层的目标则更多的是面向需求，考虑用户体验。同时，分层解耦不仅可以做到数据的模型复用，可以降低数据处理各阶段的耦合程度，同时有助于评估、分析及追踪数据在不同处理阶段所消耗的系统资源，并调整优化硬件配置。

- **层级间禁止逆向依赖：**数据的流转不应当出现循环依赖的情况。所有的数据都需要有最终认可的信任源头，逆向依赖不仅不能保证数据质量，同时对后续的数据运营也会有很大的影响。

- **模型的可扩展：**在建模阶段，我们需要基于业务而非需求进行建模，需求的变化是远远快于业务的，因此，我们需要确保模型尽可能多的满足所有需求，但结合落地成本，我们需要确保在需求变化时，可以快速生成模型及其对应的初始化数据，已满足需求的快速响应。

- **历史业务场景可追溯：**随着业务的发展，业务场景会随之发生改变，那么就需要保证模型满足对于历史业务数据的追溯，以确保满足各种需求。

数据接入

数据接入，即为了满足数据统计、分析和挖掘的需要，搜集和获取各种数据的过程。数据接入作为数据应用的源头，目的是自动化、规模化地从各个数据源去采集收集业务数据。在数据接入的前期阶段，需要从以下几个视角考虑：

- **数据质量探查：**在接入前期，要对即将接入的数据基于业务输入进行质量探查，这样做一方面可以在数据接入前期了解数据的整体质量情况，另一方面可以反过来验证对于业务的理解是否完全正确，以防前期梳理业务对于一些特殊场景的疏漏导致的后续数据质量问题。

- **数据时效与频率保障：**需要基于需求的时效性诉求结合数据源来确定数据接入的频率，结合实际数据量级综合评估如何满足业务诉求。

- **数据保留时长：**基于法律法规和企业的监管要求，来决定待接入数据需要保存的时长。通常，规范的数据生命周期管理，可以提高数据的整体管理水平，同时满足监管要求。

- **数据安全规范：**数据接入阶段，不仅需要考虑数据的传输和存储安全，同时需要了解待接入数据是否包含 PII 数据以及对其是否需要进行特殊处理以满足监管要求。

在前期阶段准备完成后，接下来就需要基于以下原则来确保数据的时效性和完整性。

- **监控业务系统变化的能力：**当上游系统发生变化时，需要及时识别并告警，变化情况包括但不限于网络不稳定、业务系统宕机、采集通道异常、数据格式改变等。不论是基础设施或是数据层面，数据接入都应当有识别变化的能力，只有提前发现业务系统变化，才能及时做出后续应对措施。

- **保证数据完整性：**面对不同的数据源类型和数据格式，采集方式也会多种多样，因此采集过程中难免会遇到各种问题。当问题修复后，需要确保待采集的数据依然可以被正常有序地采集，从而确保后续处理数据的完整性。

- **在存储和监管满足的前提下，尽量保存每一次的快照：**从数据源头获取的数据，无论其是全量还是增量更新，在采集层尽可能地确保数据都被完整记录下来，这样一方面可以做到数据的可追溯性，同时，在数据建模阶段，为了成本考虑，通常不会对所有的需求进行建模，那么当有少量需求不满足时，可以通过快照数据快速初始化模型数据，以满足快速响应需求的要求。

- **不进行业务逻辑处理：**在数据接入阶段，通常目标是尽可能地确保所采集的数据格式、类型等和数据源保持一致，而无需对数据进行业务逻辑处理。

图：数据接入的阶段和关注角度

数据准备	数据接入
数据质量探查	监控业务系统变化的能力
数据失效与频率保障	保证数据完整性
数据保留时长	在存储和监管满足的前提下，尽量保存每一次的快照
数据安全规范	不进行业务逻辑处理

数据处理

数据处理，即为了满足数据加工、计算过程的需要，在确保顺利完成数据计算时能够对出现的异常情况进行捕捉、再处理、预警的过程。在数据处理阶段，遇到的两大挑战分别是如何保证数据质量和如何降低运维成本，那么我们通过以下原则来解决：

- **满足 ETL 的幂等性：**通常对幂等性的定义是相同的参数重复执行得到相同的结果。ETL 的幂等性就要求 ETL 可以被重复多次执行，且不会影响最终的计算结果。在面对复杂的数据流时，数据处理过程中的异常或日常运维需求都意味着 ETL 可能会随时停止、随时启动，那么如何在 ETL 重复多次执行的情况下确保数据的准确性和一致性就极为关键。满足 ETL 幂等性的核心逻辑在于处理数据阶段待处理批次的数据队列清晰有序且可控，同时对于所涉及数据要满足业务依赖。从运维视角看，运维人员可以在不同需求场景下对 ETL 进行手动触发，而不用担心是否会影响数据的准确性，从而可以在保证数据质量的前提下降低运维成本。从设计视角来看，则是要将调度依赖和数据依赖进行解耦，这样就能确保调度层面的异常不会影响到数据本身。从混沌工程的原则看，能确保在满足数据质量的前提下，降低计算资源浪费。

- **日志分级分类解耦，一站式查看：**数据处理会涉及到任务调度服务、资源调度管理、计算、存储等多种技术组件，而在数据处理阶段，每一个组件的异常都会导致数据处理的失败，那么在定位问题时就需要去各个组件中查看问题的根源，这就导致了运维成本大大增加。因此需要将日志进行分类解耦，资源层面、调度层面、计算层面、数据层面等不同数据问题进行分类，可以帮助我们更便捷地开展运维工作。同时，对数据的错误也进行了分级，在数据处理阶段，对于异常数据不能进行一刀切的方式处理，而应当根据业务来决定异常数据的错误级别，哪些数据可以流入数据平台，哪些需要被清理掉，在数据处理阶段需要明确定义各类数据错误的处理规范。除此之外，推荐有统一的门户进行日志查询，可以更方便的进行运维管理。

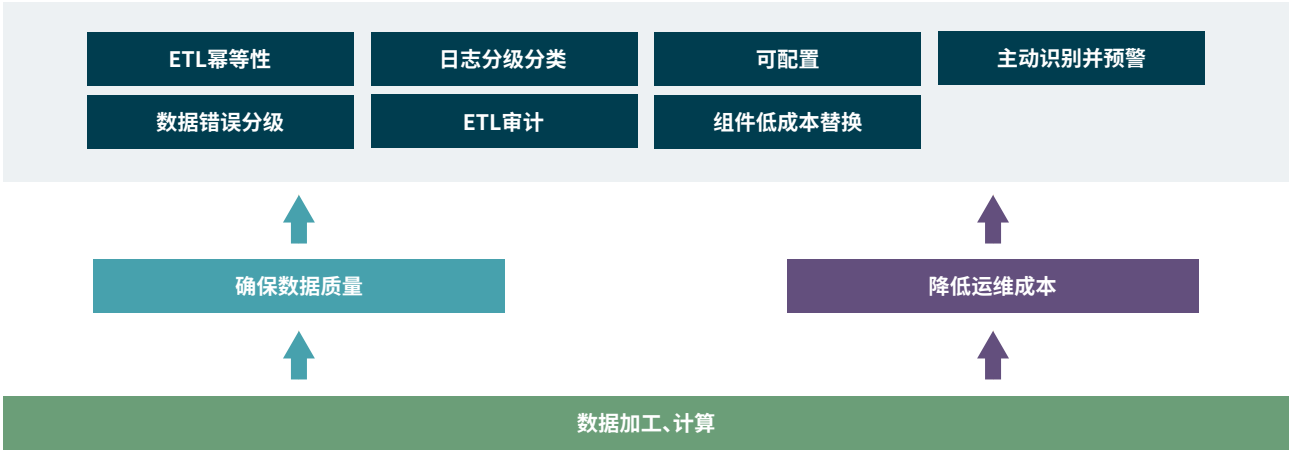
- **主动识别数据质量问题并预警：**数据问题不可避免，那么在数据出现异常后，如何识别并快速做出响应，而不是由数据使用方发现异常，这样不仅会导致平台的数据信誉度下降，异常数据流入到下游会严重影响到业务。因此，在数据处理阶段，就需要对异常数据识别并预警，一方面可以提前预知使用数据的异常可能，另一方面，可以推动业务系统或流程的完善。

- **ETL 可观测：**面对多业务数据、多调度批次的数据处理场景，出现不同级别的异常后就需要对数据流进行定位，因此 ETL 的审计可以提高 ETL 对 IT 及业务用户的友好程度，降低确认日常数据处理任务结果的复杂性，并有助于提高用户追踪异常及异常数据的效率。

- **组件低成本替换：**随着数据平台的发展，会出现某些组件已经不适用于当前场景的情况，因此需要避免对于组件的过度依赖以防止 ETL 的部署迁移或组件替换时，由于组件的高度耦合导致成本变高。

- **可配置原则：**在进行 ETL 开发阶段，尽可能减少对于常规数据处理的代码重复开发，一方面可以节省开发成本，另一方面可以提高数据质量，避免面对不同场景的数据处理考虑疏忽导致的数据问题。同时，对于沉淀框架的 ETL 开发工具，可以做到低运维、低开发成本。最后，配置化的代码可以做到统一管理，方便后续数据运营。

图：数据处理的关键工作



数据测试

测试，是信息系统交付必不可少的环节，是为了确保信息系统的正确性、完整性和安全性等而进行的一系列操作的过程，其最终目标是为了保证信息系统的品质。数据工程作为数据信息系统落地的过程，测试同样尤为重要。

- **重新定义测试金字塔**

在传统软件开发过程中，测试金字塔理论已经成为经典测试理论指导着测试的推进。其最早由 Mike Cohn 于 2009 年的著作《Succeeding with Agile：Software Development using Scrum》提出，其表现形式为一个三层金字塔结构，从下到上依次为 Unit Test（单元测试）、Integration Test（集成测试）、End to End Test（端到端测试），下层代表测试投入多，上层代表测试投入少。

测试金字塔的内核为：在一定的测试资源投入条件下，通过成本较低的单元测试扩大覆盖比率，而成本较高的端到端测试则要尽量覆盖主业务流程，辅以集成测试保证系统之间稳定调用。

在数据工程领域，测试金字塔内核仍然适用，我们将测试金字塔重新定义为：

- **单元测试为基础确保最小逻辑的准确。**其涵盖两方面：一、数据工程的基础是 ETL，大部分数据工程均会有一些工具来自动生成 ETL，而 ETL 自动生成代码，就必然少不了单元测试。二、有了 ETL 之后，ETL 内部仍然是由多个功能活方法组合而成，针对 ETL 内部方法的单元测试仍然不可或缺。由于单元测试相对独立，编码成本较低，可以以小的代价运行。并且 ETL 为数据工程事实上的基本单位，对其进行的单元测试可以覆盖大部分细粒度的逻辑。

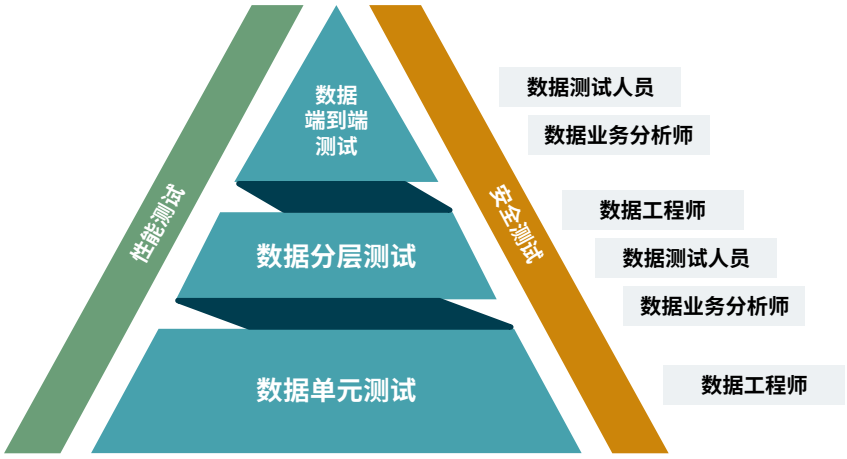
- **分层测试确保单个模型的数据质量。**在数据工程当中，为了快速响应变化、提高重复利用率以及减少性能瓶颈，大部分的数据架构是纵向分层的架构，而不同层次有不同的数据处理逻辑，那么就需要先对每一层先进行独立测试验证，再重点测试层与层之间的集成与功能。测试关注：元数据验证、数据值、处理逻辑与处理性能等。在保证每层数据、逻辑正确的情况下，才能为更高层次的功能与数据质量提供保证。

- **数据端到端测试确保交付需求的质量。**端到端测试是从数据源到最终结果的验证过程。覆盖了数据全链路层与层之间的耦合逻辑。一般而言，从数据源头到最终数据应用链路很长，计算资源消耗也比较高，进行端到端测试的方法一般是通过构建源数据，直接对比处理末端或应用端数据结果是否符合预期。数据端到端测试虽然可以从最终结果上校验功能，但其存在成本较高，数据用例构造复杂度较高、发现 Bug 定位困难、运行时间超长等弊端，所以这层一般更多的是进行 happy path 的验证与端到端性能测试，不会大范围覆盖所有分支逻辑。

- **安全与性能测试。**测试金字塔一般用来当做面向功能的测试策略。除了以上讲到的在金字塔内部的多层测试，在数据领域，由于数据量巨大以及数据往往会涉及到各种机密与隐私，所以数据安全测试、性能测试同样很重要。数据安全一般会根据具体项目情况涉及不同的测试策略，详情可参阅数据安全篇章。而数据性能则是另一个比较重要的点，一般的步骤为：预计数据量级，构造数据、准备生产仿真环境、准备测试用例、产出性能测试报告、分析与改造等。

- **人员与能力标准。**数据工程测试金字塔从下到上技术细节逐渐减少，业务含义逐渐增多，通常来讲，底层 ETL 测试主要由数据开发人员负责。中部数据分层测试由于包含对数据模型的验证，需要有一定业务理解能力的人员参与测试用例的制定，一般由数据测试、数据业务分析师以及数据工程师共同参与。而顶层的测试用例由于很少涉及编码细节，其测试基本可以由数据分析师和数据测试共同完成。

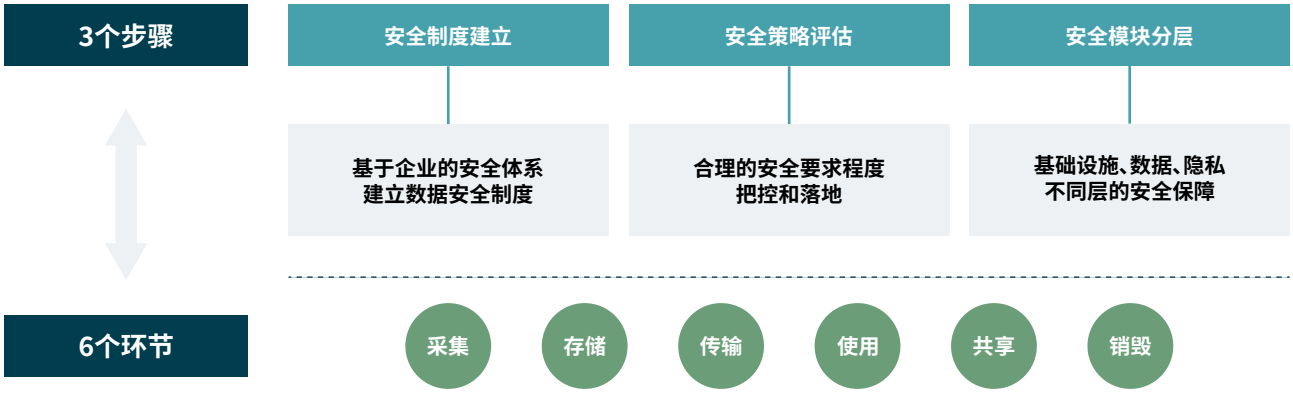
图：数据测试工作全景



数据安全

数据既是生产要素，也是企业的重要资产，如何保障数据的安全就显得尤为重要。安全不是简单的规章制度，需要系统性的构建企业数据安全体系。

图：数据安全的三个步骤和六个环节



- **数据安全要融入企业安全体系中：**数据产生于业务，又是业务的一部分，故而数据安全应与企业资产安全一一对应。什么安全级别的业务，其对应的数据也至少需要有相应的安全级别。例如，产品研发企业，由于新产品是其核心竞争力，有着最高安全级别，相应的，其新产品对应的研发数据也应该做到最高安全级别，而其营销数据的安全级别可能就不需要那么高的要求。

- **数据安全是变化的，可演进的：**由于大部分企业对数据的管理和应用还属于摸索阶段，还处于数据能跑能用就行的状态，对数据安全还没有明确统一的认知。早期建立的数据安全体系不一定适应现在的数据安全要求。随着技术壁垒的打破、新技术的更新迭代，数据安全策略需要持续的提升和优化。但是对于企业来讲，精力和资源都是有限的，因此我们需要把有限的精力和资源投入到合适的地方中去，这也就意味着我们需要对安全要求程度的把控和预期要有一定的控制。所以在这里提出结合我们的经验和实践站在安全视角下的考察维度。

- **数据安全实施细则：**数据安全说到底是信息系统的一部分，是属于跨功能需求。在具体数据安全落地过程中，我们可以分层次的来构建数据安全，从技术设施到功能模块全方位构建企业数据安全体系。一般我们将数据安全落地分为以下三个层次：

1. **基础设施安全：**基础设施安全主要针对基础设施涉及到的安全隐患，包含数据传输、数据存储、数据计算、管理平台等。
    - **数据传输：**需要考虑接口是否鉴权、传输协议是否安全、传输管道是否加密等因素。
    - **数据存储：**需要考虑文件系统是否加密，备份与容灾机制是否健全，存储介质是否可访问等因素。

- **数据计算：**需要考虑是否有身份认证，密钥信息是否有妥善保管机制，计算过程中是否出现明文密钥信息等。
- **管理平台：**需要考虑操作系统是否及时安装补丁，配置管理是否安全，托管平台是够安全等等。

2. **数据安全：**数据安全主要针对数据在访问、使用过程中以及过程后可能出现的各种安全问题。包含：
   - **数据加密：**在各种复杂计算机系统中，数据加密能够有效降低数据泄露带来的风险，即使数据被泄露，在没有密钥的情况下也很难从数据中获取有效价值。数据加密一般分为对称加密和非对称加密，具体算法可由具体情况而定。并且，定期轮转加密密钥也能有效降低数据泄露风险。
   - **数据隔离：**企业数据平台往往会整合众多业务系统数据，给不同业务域的人员使用，数据隔离能够有效划分数据界限，理清数据管理权限，帮助更好的管理数据资产。
   - **数据访问控制：**在数据隔离的基础上，针对不同角色的操作用户，划分不同权限，保证对数据权限的严格控制，做到每种角色对所需数据权限最小化原则，并提供权限申请功能，将数据权限管理纳入到流程之中，充分做到数据请求合理合规。
   - **数据溯源追踪：**在海量数据汇聚在数据平台的背景下，能够对数据链路追踪与溯源，在发现数据安全问题的情况下能够有迹可循，快速确定影响范围并及时补救，防止危机扩大。
   - **数据管理：**海量数据必然会产生大量的元数据，有效的元数据管理能够保证数据安全的有序推进。
   - **数据销毁：**在某些特殊情况下，对于敏感信息、机密信息，需要提供有效的数据销毁机制，来保证机密信息不被窃取，如企业内监管要求对部分数据的生命周期有对应的时间要求等。
   - **监控与审计：**数据既然作为企业资产，那么所有数据的读取、操作都需要记录相关操作记录，既可以用来分析企业数据安全状况，有效发现程序后门，还能帮助分析性能，更重要的是，对于不合规或危险操作，能够及时预警，将数据安全问题做到早发现、早解决。

3. **隐私信息保护：**针对数据中可能包含的各种隐私数据，要避免其出现泄露，尤其针对一些公共暴露的数据 API 更要关注隐私信息安全，包含：
   - **去标识化：**数据去标识化是从数据中移除标识信息的过程。标识一般分为直接标识符合准标识符，直接标识符是指能够直接定位数据主体的数据，包括姓名、住址、身份证号、电话号码等；准标识符不能直接定位数据主体，但可以通过组合识别出数据主体，比如邮编、公司、生日、性别等。

能力复用与保障

数据工程落地过程中，不仅会沉淀数据资产，亦会沉淀 IT 资产。模型的复用、工具的沉淀、平台的搭建，均是数据工程落地过程中能力复用的具体表现。

数据工程与应用工程的最大区别在于，软件工程实现的是业务流程，而数据工程实现的是信息与数据流程。信息与数据流程天然通用一套范式，我们可以尽可能多地将通用能力抽离出来，以工具、平台的方式沉淀下来，从而加速基础设施的演进与发展、加速新功能的孵化、提升开发与运营效率。

我们将能力复用分为三个层次，即：

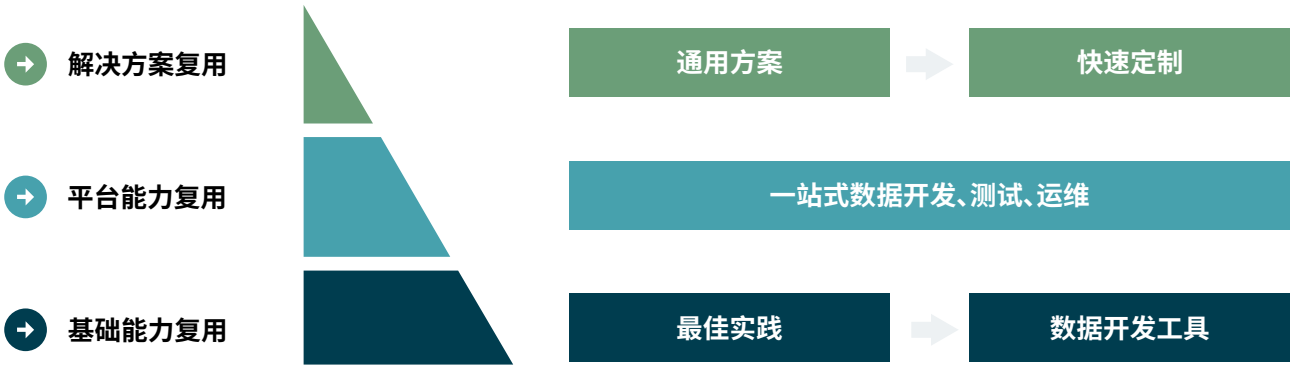
- **基础能力复用：**对于数据工程开发过程中的最佳实践，我们可以将其沉淀为数据开发工具，例如数据运维、中间表生成、ETL 自动生成、监控告警等等。不同工具灵活组合，又由于不同工具可以提供灵活配置，可以满足数据开发工程师、数据分析师、数据运维工程师以及数据测试工程师等多种角色的不同诉求。故数据工具需要满足可配置性、低耦合等特性。

- **平台能力复用：**一般而言，数据工程落地的形式都会是各种企业内部数据平台。数据平台的特性是各个功能模块相互配合，可以提供一站式数据开发、测试、运维功能。从而降低数据团队运维成本，提高生产效率。

- **解决方案复用：**对于新业务，如果已经构建了其所属共性业务的解决方案，则可以通过调整方案进行快速定制。方法是：基于解决方案的通用流程制定新流程，罗列共性模块与特性模块，在复用数据平台的基础上，挑选合适的基础能力，快速实现配置与开发。

本为中提到的复用是对于能力的可复用性，并非某个具体的产品或者是方案，正因如此能力复用才能作为企业数据工程能力规模化推广和应用的基础。

图：数据工程能力复用与保障



数据持续运营

在数据汇聚、整合完成后，还需要对数据进行运营，以满足数据可以被更便捷、更安全、更稳定地使用，更好地反哺业务，产生业务价值。数据运营的目的是要形成企业看数据、用数据、将数据作为沟通语言和工具的“数据文化”，数据只有容易被发现，才有产生价值的可能性。那么针对不同的数据消费角色，数据的展现形式也应该多样化，比如针对数据分析师，数据资产目录可以很方便的帮助他们找到想要的数据，而针对业务决策人员，为了更科学系统的查看指标辅助决策，数据集市则更适合他们。对企业数据资产的结构化描述、存储、搜索、管理的系统，包括元数据的搜索、浏览数据样本结构、拉通数据全貌、快速发现、定位数据服务以及数据服务的所有者等功能。那么，数据运营包含：

- **持续更新与迭代的数据资产：**数据是企业的资产，能够为企业创造价值，但是企业的业务并非一成不变、企业在价值的度量也并非一直相同。因此数据资产在前期被定义出来之后，我们需要通过后续的持续维护才能够确保数据资产的有效性。数据资产目录是数据资产盘点以及后续维护的一个载体。从概念上来看，数据资产目录将业务信息和技术信息进行关联，并提供给不同角色的数据消费者。例如：可以告诉业务人员当前都有哪些可用的业务信息、指标信息，也可以告诉技术人员，这些信息分别分布在哪个目录、哪个表等。通常，数据资产目录的业务元数据要包括主题域的分组、主题域、业务对象、逻辑数据实体、属性信息。技术元数据信息包括物理数据库、schema、表、字段。这里还需要满足一个原则：数据源头的唯一、数据所有者的唯一。

- **低成本与人工干预的数据运维：**在进行数字化转型中，会涉及比较多的组件。包括一些分布式的存储计算引擎，涉及多个数据源头，涉及多种临时数据诉求。因此每种类型都需要考虑其运维，同时，运维工作从软件的生命周期看往往占了大部分时间。从后续维护的视角来看，数据运维是重要的组成部分，也是工作量的体现部分。我们通常会将运维分为几大类。基础设施的运维、调度运维、数据运维、安全运维和其他运维。

  基础设施运维包括所涉及到的所有底层技术组件，如 CPU、内存、网络、消息队列、存储 / 计算引擎等，需要考虑大规模集群下的管理，环境隔离，容灾备份等，需要考虑集群的预警和审计，确保基础设施的稳定和可追溯。

  调度运维是指在 ETL 出现异常后的日常处理，通常需要满足日志分级分类、ETL 幂等性、预警的时效性和异常处理流程闭环，可以做到快速定位问题，快速处理解决问题且保证数据的准确性。

  数据运维通常是在数据源头发生数据异常或者数据变更的运维或一些临时的数据需求，当数据异常时，需要明确数据 owner。对于一些临时需求，如需要重刷历史某天的数据，需要做到快速相应，如可以以最少的资源满足需求，同时可以避免一些数据问题，如旧数据覆盖新数据等场景。

  安全运维包括底层基础设施的安全运维、ETL 和数据层面的安全运维。如对于基础设施的安全监控、代码的安全规范扫描、依赖代码升级、密钥信息的安全处理、机密信息的加密处理、PII 数据的脱敏处理和权限控制等。在接收到异常安全监控告警后，需要及时处理对应的安全事故。其他运维主要包括在开发部署上线阶段的运维工作等。

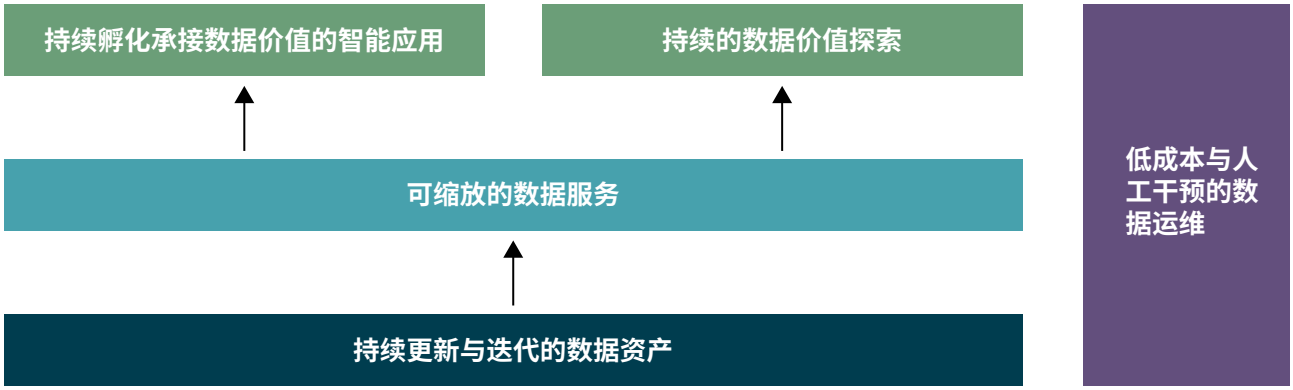
对于数据运维工作是否优秀的很重要的评判标准，可以从数据运维工作上投入的工作量有多少、在数据运维工作中需要人工干预的环节有多少，是数据工程落地实施是否优秀等几个视角来衡量。

- **可伸缩的数据服务：**数据服务作为对外提供数据的重要方式，他的目的是通过标准化数据服务将可信、易用的数据集提供出去，支撑业务的开展。但是业务是变化的，业务对于数据的使用也就是变化的，我们定义好的数据服务也是需要跟着变化的，这里就要看数据服务在被使用的频率来评估是否要对数据服务进行扩缩容；对数据服务使用的正常和异常进行监控。

- **持续的数据价值探索：**持续的数据价值探索有两个挑战，第一是如何持续的发掘有价值的业务场景；第二是如何能高效便捷地进行数据价值探索。在前面愿景对齐讲到，需要结合现有的数据和技术手段来创新性的探索业务场景，但是业务场景的解决方案和价值体现并不是一成不变的，仍然需要持续迭代。除此之外，需要提供数据自服务实验室，可以让业务人员通过可视化的方式结合自身对业务的理解来敏捷高效地探索数据，从而更大的发挥数据价值。

- **持续孵化承接数据价值的智能应用：**智能应用作为数据使用和产生价值过程中重要的载体，需要根据业务诉求以及对于数据价值的挖掘持续的探索和演进新的智能应用，并通过智能应用的构建和演进来应对新的市场和大环境带来的挑战、改善用户体验。

图：数据持续运营



数据工程能力建设

数据研发能力建设

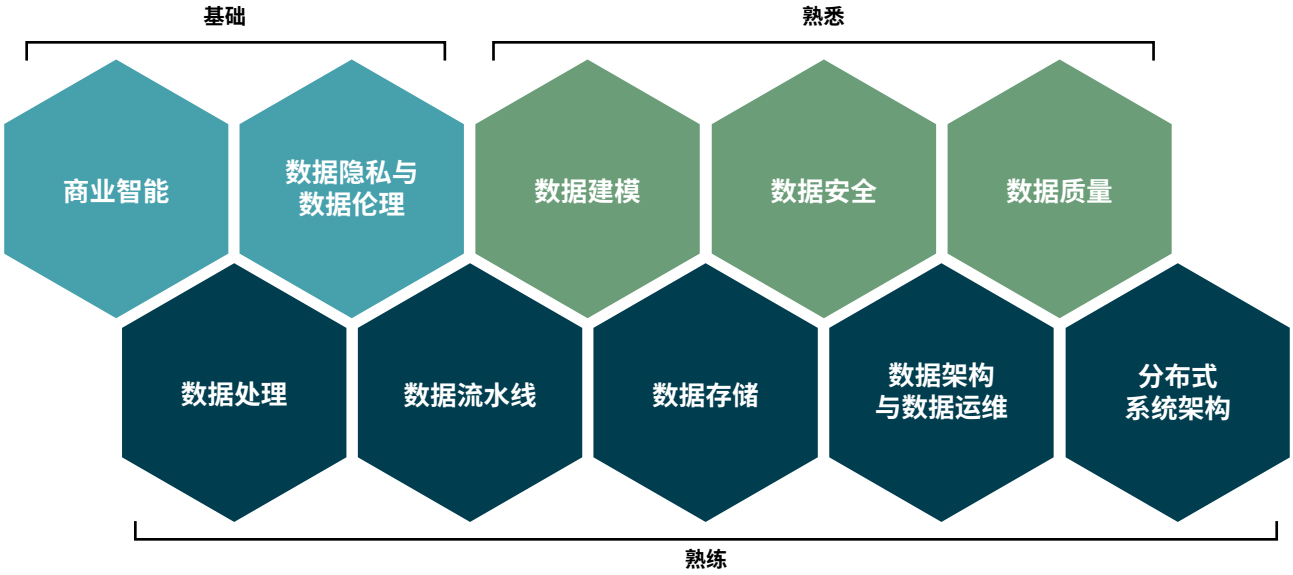
数据工程实现与落地过程中涉及到方方面面的工作，从确认需求到后期运营；从质量管控到安全保障；从设计到实施等多个维度。意味着数据工程落地过程中需要有跨部门的合作；企业的业务与数据融合后的流程管控；对技术能力的更高的要求。这也对企业落地数据工程带来更大的挑战，挑战来源于数据工程落地的时候并非单纯的技术问题，而是技术、数据、业务相融合后的结合体。因此数据工程落地能力有以下几个方面的诉求：

- **数据战略制定与数据思维推广。**数据作为新一代的生产资料，明确企业的数据战略规划有助于在企业内部建立数据能力的基础。

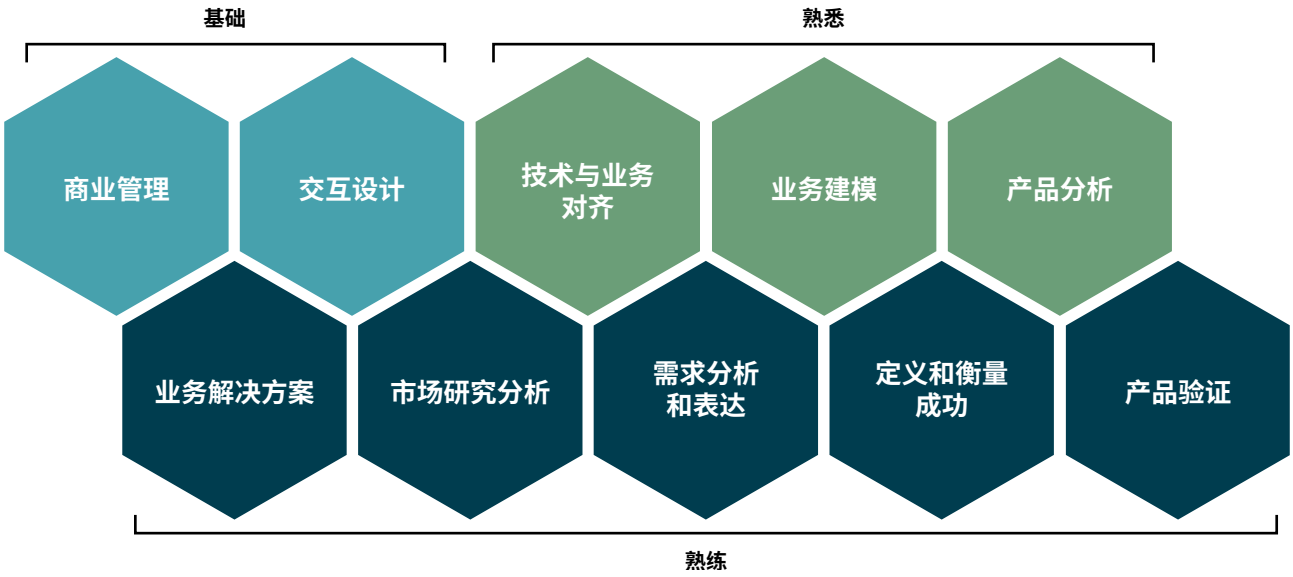
- **数据工程能力定位。**数据工程能力涉及到愿景对齐、落地实施、持续运营三个大的方向，在实际操作过程中也很可能会贯穿企业数据与业务部门，因此数据能力应该以中心化的方式还是以去中心化的方式需要结合企业内部实际情况来进行评估，但是企业需要有明确的数据能力沉淀方向。

- **数据工程人员培养。**数据工程的落地，归根结底还是需要由人来完成。构建企业自身的人员能力培养机制、搭建企业人员数据能力提升通道是数据工程能力持续迭代的重要保障，如下图所示的数据工程师能力模型，企业需要明确自身发展路径上的数据工程能力诉求，以便更好的寻找和培养数据工程人才。

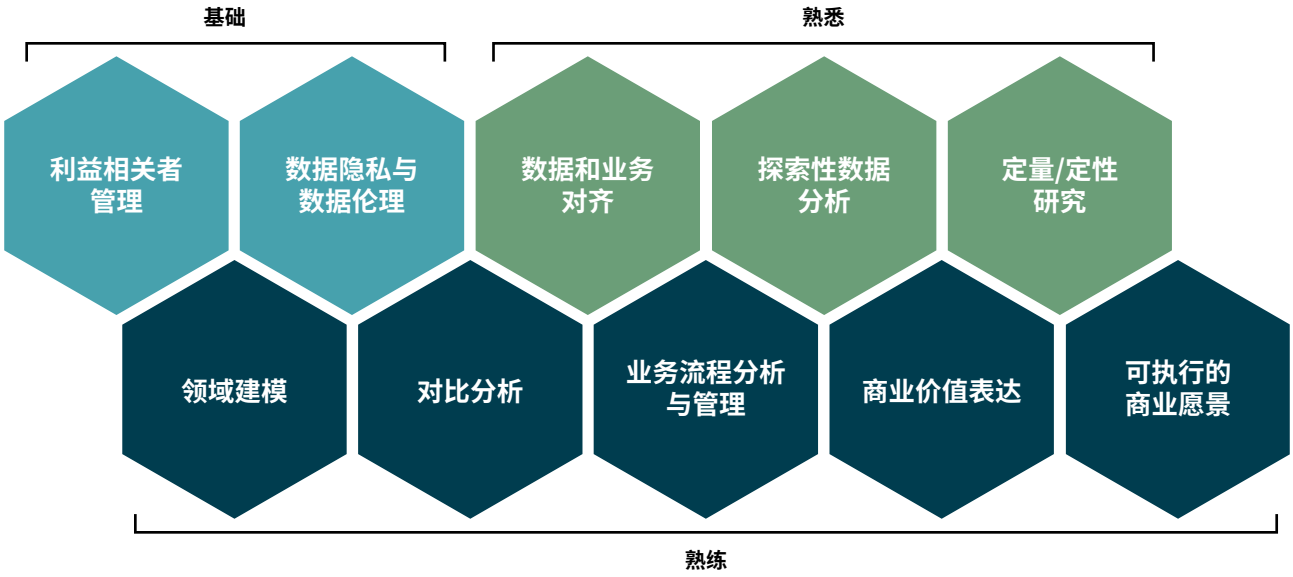
图：数据工程师能力诉求样例



图：数据产品经理能力诉求样例



图：数据业务分析师能力诉求样例



- **数据工程的跨部门合作。**数据工程落地过程需要业务、数据、技术相融合才能更好的体现出业务价值，那么也就意味着企业需要有一套跨部门协作的流程和机制。

- **数据工程知识沉淀。**数据工程能力的规模化前提在于能够快速的复制，并将数据工程能力落地，这对企业的知识沉淀提出了新的挑战，即当人员流动时、外部环境变化时，企业是否能够快速的沉淀、积累相关知识，并与上述人员培养机制相结合，完成知识到应用的过程。

因此企业需要结合自身的诉求和现状制定合适的数据研发能力建设策略，确定数据工程能力需要落在哪里；围绕数据战略的数据思维推广；提升落地工作中可复用的内容来节省成本；构建快速响应变化的流程和机制；沉淀知识为组织赋能，最终实现数据工程的落地，持续为企业带来更多价值。

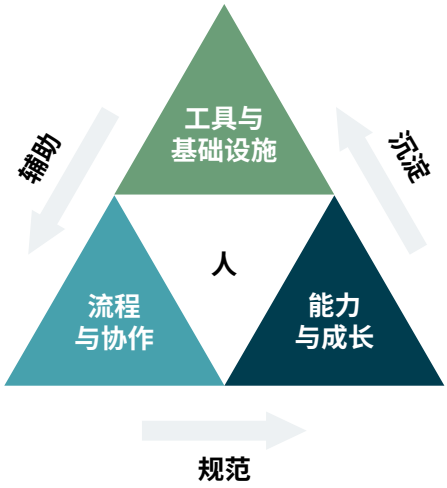
数据工程成熟度评估

通过科学的数据工程成熟度评估体系，可以帮助企业识别数据工程级别，明确优化方向。

在明确了数据工程落地所需的能力之后，如何以低成本、高效率将数据工程进行落地，就成为了企业需要关注的重点，也侧面反映了企业研发管理能力的强弱。而在数据工程落地之后评估落地质量，则要构建企业数据成熟度评估体系。

数据工程就是企业通过数据研发过程交付业务价值的过程，在这个过程中经历的时间越短、消耗的资源越少、交付的质量越高、后期扩展性以及可维护性越高，则我们可以说数据工程越成熟。整个阶段可以划分为交付阶段和运营阶段，这两个阶段可以同时存在，交付阶段以**数据研发效能**为评估核心，运营阶段则以**数据持续运营**为核心。

图：数据工程落地三要素的关系



企业管理者往往会将效率与速率混为一谈，速率只关乎速度，而效率关乎资源，其可以从不同维度考量，包括：时间、人力、金钱、资产等等。所有的核心都是人，如何围绕研发团队构建一整套体系，可以描述、反馈以及提升研发效率，才是研发效能成功构建的关键。

- **基础设施与工具。**基础设施是研发团队的生命线，良好的基础设施能够极大的加速研发效率。数据工程由于技术生态繁杂，单传统批数据处理工具就有 Hadoop“动物园”数十种工具链，更不用说在机器学习领域的各种框架、算法以及相关的引用库，有可能一个引用库版本错误就可以导致整个项目无法启动，而不完善的基础设施会占用开发人员大量精力。数据工程的基础设施的终极形态应该是云服务那样的一站式、一键式环境搭建工具，并能够提供完整的 DevOps 能力。工具是在数据领域抽象出来的一系列研发能力集合。大多数企业往往容易忽略研发工具体系的长尾效应，即企业过分追求端到端、全流程、一站式，功能齐全的工具，在数据领域，一站式带来的效能提升并没有想象的高，专业的事情交给专业的工具去做，理想的形态应该是在标准化接口下的可插拔式的工具集合。

- **流程与协作。**当前大多数企业的研发效能体系建设还停留在基础设施与工具的构建上，而这只是研发效能的一部分。研发的核心是研发团队，有团队必然有协作，有协作必然有流程，流程太长会导致效率低下，而流程过少会导致管理手段失效，在软件研发领域经常提到的敏捷、SCRUM、瀑布等都是流程与协作的经典模式，选择什么样的流程与协作模式，需要根据不同企业的需求、团队规模、人员水平等共同考量。比如小规模团队，面对快速变化的业务需求，敏捷的工作方式就比较合适。而面对确定业务需求，有明确交付时间节点的企业，可能瀑布模式就是一个比较适合的模式。

- **人员能力与成长。**不同企业规模不同，需要管理的研发团队结构也不同，有的企业数据研发团队全部为自有人员，有的企业规模较大并且引入了大量的外派开发人员，并且研发团队人员能力参差不齐。研发效能一般会随着研发团队人员规模的扩大而不断降低，但由于规模的提升，研发速率也会提升。面对这种情况，我们对于研发效能的基本要求就成了如何保证研发效能基线，不至于研发效能下降太快。面对这种情况，要守住研发效能基线，就要围绕研发人员的能力做好三点：

- **首先是加快信息流动性，减少信息孤岛。**可参考措施有：构建统一文档、定期培训、定期组织分享等等，不同团队可根据自身情况量体裁衣。
- **其次是细化分工。**细化分工能够明确每个团队成员的工作，降低每个节点的能力需求。当然，细化分工并不意味着固化分工，频繁的信息流动与基于员工兴趣的定期的轮岗可以提升整个研发团队的稳定性。
- **最后是注重能力培养。**数据工程研发能力并不是一锤子买卖，数据领域技术日新月异，团队需要为成员提供途径提升能力，提升研发团队的进化能力。

- **研发效能评价体系。**评价研发效能 , 最终是通过一系列成体系化的指标来衡量的。指标需要经过精心设计 , 常见的有：
  - **速率类型指标** , 比如构建与部署速率 , 需求到落地速率等 ;
  - **质量类指标** , 比如缺陷数量、缺陷比例、测试覆盖率等 ;
  - **耗时类指标** , 比如修复流水线耗时、平均修复缺陷耗时、脚手架构建耗时等等。

  这些指标一般是从时间、质量、数量、频率等多个维度衡量研发效能 , 企业可根据自身情况构建适合自身的研发效能评价体系。

- **数据运营评价体系。**数据运营的评价指标与研发效能关注点不同，可以通过构建完整的指标体系来度量，指标体系构建亦可参见研发效能指标体系的维度，如速率类、数量类、比率类等指标，数据运营的评价主要包含 :
  - **数据运维指标**，比如系统宕机率、错误定位时间耗时、运维人员规模等 ;
  - **数据服务响应**，如是否支持可伸缩的数据服务、是否支持升级与降级、系统响应时长 ;
  - **数据持续探索**，数据获取便捷程度、数据探索便捷程度以及数据探索成果孵化能力等，而数据探索成果孵化性能、新服务，又与数据研发效能息息相关。

数据工程展望

数据工程是数字经济下确保数据价值转化的重要保障，是加速数据转化为价值的重要手段，数据工程能力应对的不仅仅是当下的挑战，更是应对未来数字经济大趋势的秘密武器。随着需要处理的数据量的增长，为了处理数据领域的各种新问题，各种新技术、新概念逐渐涌现，现代数据仓库、数据湖、湖仓一体、分布式数据架构、机器学习、数据云原生等逐一登上舞台，数据工程的发展道阻且长。

正如本白皮书引言部分提到的“数据已经成为继土地、劳动、资本、技术之后的第五大生产要素”，任何一次科技革命都会为企业、社会甚至是全球带来冲击，我们现在正处于技术革新的过程中。这既是挑战也是利好，挑战在于我们需要重新审视企业未来的发展方向、积累和沉淀新的技术能力并与实际的业务相融合，企业需要着手开始自己的数据能力布局；利好的是企业战略布局和转型的主动权在自己的手里。

国家层面已经将数字经济定义成新的经济形态，在此新的经济形态下将数据资源作为关键要素，推动整个社会的变革将数据价值转变与配套的服务能力与国家经济指标挂钩。

这也在变相的要求企业站在新一轮技术变革的过程中，加速企业转型的速度，“产业数据化，数据业务化”不再是喊喊口号，而是实实在在的转型。在此过程中，数据工程的好坏从某种程度上来讲并非是技术能力的强弱，而是对于生产要素加工、转换为价值能力的区别。因此企业需要结合自己的战略发展与自身业态模式的特征，制定数据能力的演进路线，构建企业自己的数据能力，持续完善并沉淀数据工程能力。

所以无论是站在企业内部的发展诉求还是站在企业所处的社会大环境，都在要求企业加速自己转型、完善自己的数据能力，在激烈的市场竞争过程中获得有利地位才能在未来数字经济繁荣成熟期到来之前占据有利战略发展位置。

Thoughtworks Data & AI 数据与人工智能事业部

企业希望利用其数据资产实现最大化的业务价值。但低效的战略、复杂的基础架构和陈旧的流程令达成这一目标步履维艰。如果企业的数字化转型战略中缺少数据视角的思考，那么数字化产品团队就难以借力数据服务，企业高管也无法获得优化决策所需的数据支撑。

针对这一问题，行之有效的解决方法是——用经过实践验证的数据战略来解决基础架构的复杂性问题，并确保在业务运营中有效地融合数据管理和数据分析。充分利用数据工程、数据战略和分析以及机器学习的优势，让数据创新举措持续高效地产生价值，并最终通过数据生态系统创造巨大的价值。

我们与您合作，利用我们在精益数据创新和数据科学、人工智能方面的专长，结合近 30 年的软件工程卓越经验和在金融、汽车、制造、零售等领域的积累，帮助您创建数据创新文化、基础设施、以及数据生态系统，放大创新机会，快速交付价值，实现数据驱动的企业愿景。

联系我们

电话：4008900505
邮箱：marketing-china@thoughtworks.com

Thoughtworks
服务号

Thoughtworks
洞见

关于思特沃克（Thoughtworks）

思特沃克（Thoughtworks）是一家集战略、设计和工程于一体的全球软件及技术咨询公司，致力于推动数字创新。我们在 18 个国家 / 地区设有 50 个办事处，拥有 12,000 多名员工。在过去近 30 年的时间里，我们凭借业务和技术优势帮助客户解决了众多复杂的商业和技术问题，与客户一起实现了非凡的影响。

/thoughtworks