

当前，大部分企业不再建设从源数据采集到分析应用的烟囱式系统，更倾向于数据集中采集、存储，并应用分层建设。这种方式一方面有利于应用系统的快速部署，另一方面也保证了数据的集中管理与运营，体现数据的资产、资源属性。笔者根据个人数据中台的工作实践和学习以及思考总结，撰写成本文数据中台知识体系。

## 一.数据中台是什么

### 01 定义

数据中台是一套可持续“让企业的数据用起来”的机制，是一种战略选择和组织形式，是依据企业特有的业务模式和组织架构，通过有形的产品和实施方法论支撑，构建的一套持续不断把数据变成资产并服务于业务的机制

数据中台是处于业务前台和技术后台的中间层，是对业务提供的数据能力的抽象和共享的过程，数据中台通过将企业的数据变成数据资产，并提供数据能力组件和运行机制，形成聚合数据接入、集成、清洗加工、建模处理、挖掘分析，并以共享服务的方式将数据提供给业务端使用，从而与业务产生联动，而后结合业务系统的数据生产能力，最终构建数据生产>消费>再生的闭环，通过这样持续使用数据、产生智能、反哺业务从而实现数据变现的系统 and 机制。



### 02 本质

数据中台服务于数字化转型，而企业数字化转型的终局是传统业务变成数字化业务，数字化业务的本质就是以数据作为新生产要素进行加工，构建以数据作为主要存在形式的产品，产生商业价值的业务模型。

因此数据中台的本质更像一种企业架构，是一套互联网技术和行业特性，在企业发展的不确定性中，寻找确定性，并且持续沉淀和提炼企业核心能力，最终支持企业快速、高效、低成本进行业务创新和增强的企业架构。

### 03 数据中台、数仓、大数据平台的区别

#### 1) 数据中台VS数据仓库

数据仓库的主要场景是支持管理决策和业务分析，而数据中台则是将数据服务化之后提供给业务系统，目

标是将数据能力渗透到各个业务环节，不限于决策分析类场景。数据中台持续不断地将数据进行资产化、价值化并应用到业务，而且关注数据价值的运营。

数据中台建设包含数据体系建设，也就是数据中台包含数据仓库的完整内容，数据中台将企业数据仓库建设的投入价值进行最大化，以加快数据赋能业务的速度，为业务提供速度更快、更多样的数据服务。数据中台也可以将已建好的数据仓库当成数据源，对接已有数据建设成果，避免重复建设。当然也可以基于数据中台提供的能力，通过汇聚、加工、治理各类数据源，构建全新的离线或实时数据仓库。

## 2) 数据中台vs大数据平台

大数据基础能力层：Hadoop、Spark、Hive、HBase、Flume、Sqoop、Kafka、Elasticsearch等。在大数据组件上搭建的ETL流水线，包括数据分析、机器学习程序。数据治理系统。数据仓库系统。数据可视化系统。

数据中台应该是大数据平台的一个超集。在大数据平台的基础之上，数据中台还应该提供下面的系统功能：

- 全局的数据应用资产管理
- 全局的数据治理机制
- 自助的、多租户的数据应用开发及发布
- 数据应用运维
- 数据应用集成
- 数据即服务，模型即服务
- 数据能力共享管理
- 完善的运营指标

## 二.数据中台核心能力

数据中台需要具备数据**汇聚整合**、**数据提纯加工**、**数据服务可视化**、**数据价值变现** 4个核心能力，让员工、客户、伙伴能够方便地应用数据。

### 1、汇聚整合（数据治理-数据整合和管理能力）

数据丰富和完善：多样的数据源进行合并和完善

- 管理易用：可视化任务配置、丰富的监控管理功能
- 数据集成运营：数据接入、转换、写入或缓存内部来源的各来源数据
- 数据目录与治理：用户可以方便定位所需数据，理解数据（技术/业务治理）
- 数据安全：确保数据的访问权限
- 数据可用：用户可简便、可扩展的访问异构数据，可用性和易用性高
- 部署灵活：本地、公有云、私有云等多种部署方式

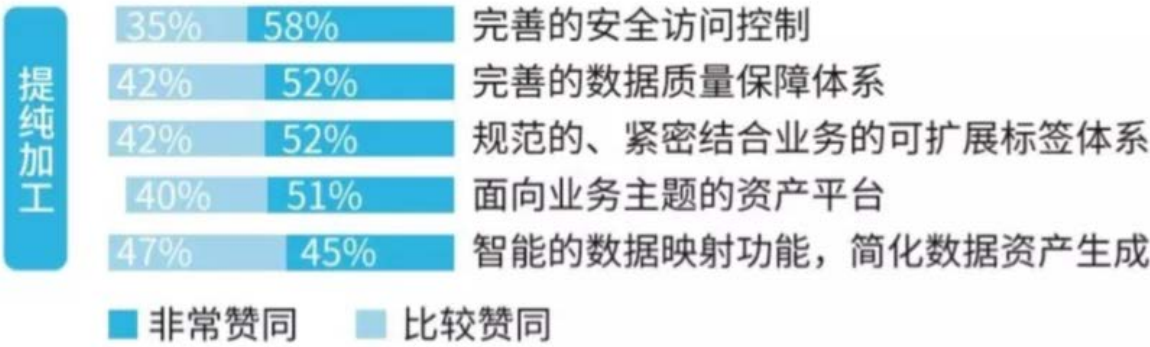
企业看重的数据整合和管理能力（非常认同%）



2、提纯加工（数据资产化——数据提炼与分析加工能力）

- 完善的安全访问控制
- 完善的数据质量保障体系
- 规范的、紧密结合业务的可扩展的标签体系
- 面向业务主题的资产平台
- 智能的数据映射能力，简化数据资产生成

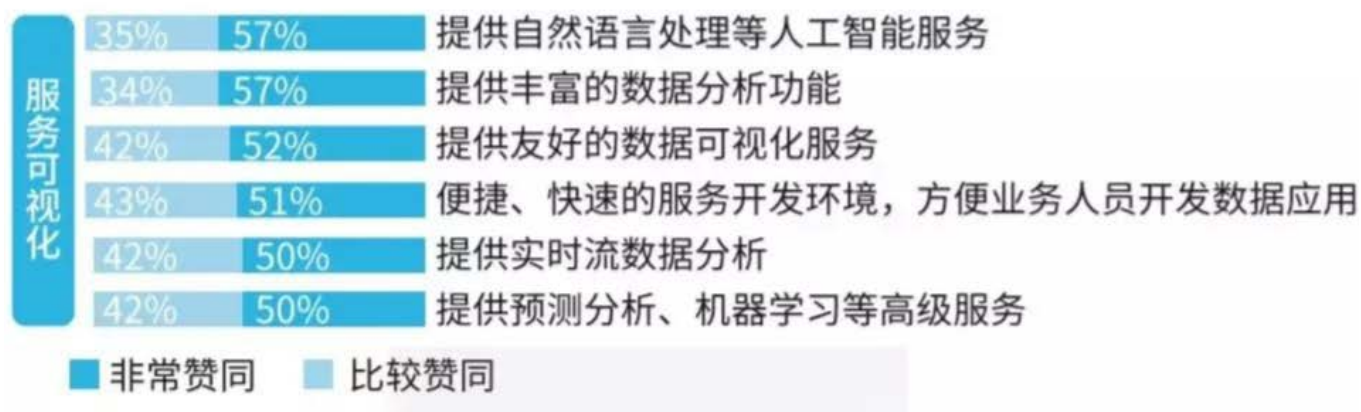
企业看重的数据提炼和分析加工能力（非常认同%）



3、服务可视化（数据资产服务化能力）

- 提供自然语言等人工智能服务
- 提供丰富的数据分析功能
- 提供友好的数据可视化服务
- 便捷、快速的服务开发环境，方便业务人员开发数据应用
- 提供实时流数据分析
- 提供预测分析、机器学习等高级服务

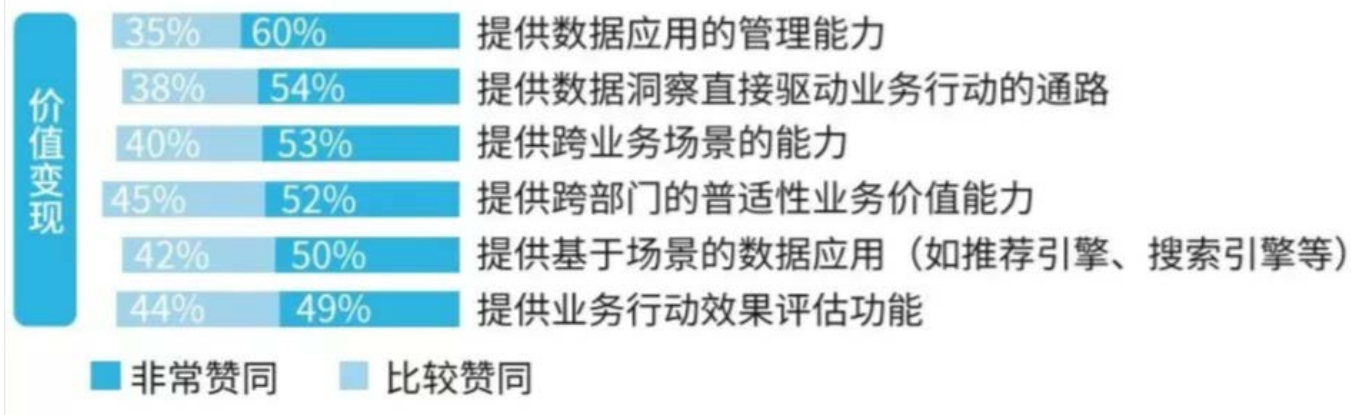
### 企业看重的资产服务化能力（非常认同%）



#### 4、价值变现

- 提供数据应用的管理能力
- 提供数据洞察直接驱动业务行动的通路
- 提供跨行业务场景的能力
- 提供跨部门的普适性业务价值能力
- 提供基于场景的数据应用
- 提供业务行动效果评估功能

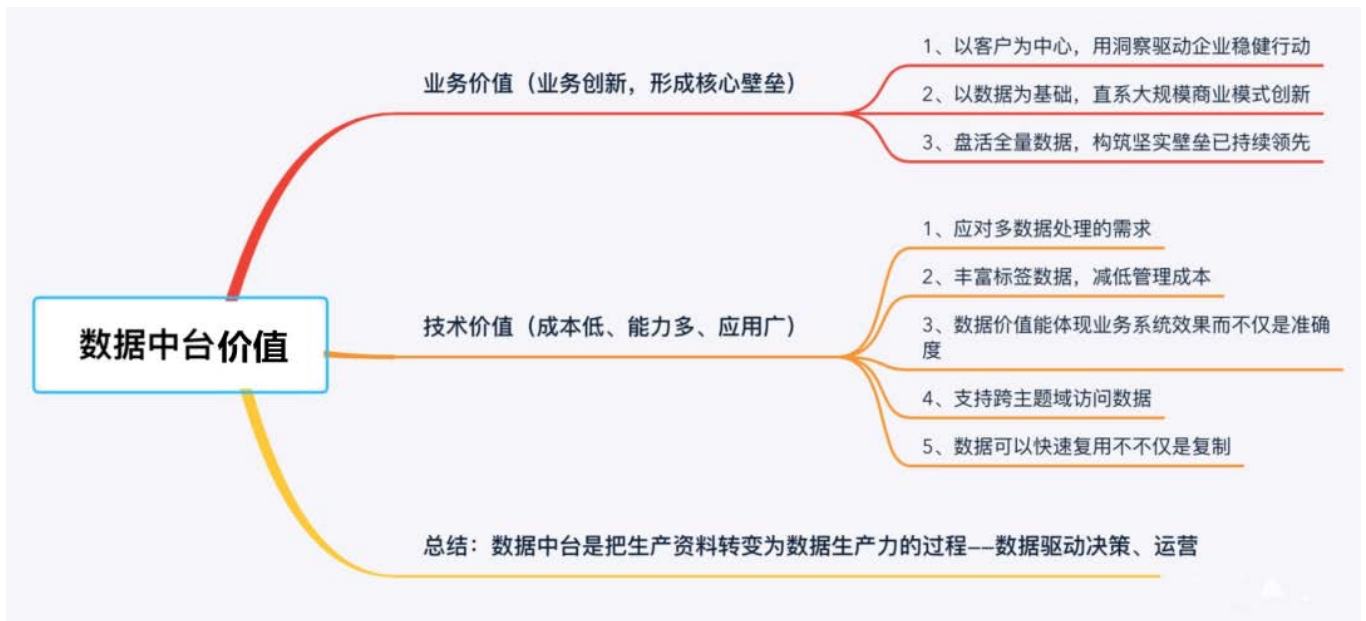
### 企业看重的业务价值变现的能力（非常认同%）



### 三.数据中台的价值

数据中台价值如下：





## 1) 业务价值（业务创新，形成核心壁垒）

- 1、以客户为中心，用洞察驱动企业稳健行动
- 2、以数据为基础，直系大规模商业模式创新
- 3、盘活存量数据，构筑坚实壁垒已持续领先

## 2) 技术价值（成本低、能力多、应用广）

- 1、应对多数据处理的需求
- 2、丰富标签数据，减低管理成本
- 3、数据价值能体现业务系统效果而不仅是准确度
- 4、支持跨主题域访问数据
- 5、数据可以快速复用、不仅是复制

总结：数据中台是把业务生产资料转变为数据生产力，同时数据生产力反哺业务，不断迭代循环的闭环过程——数据驱动决策、运营

## 四.数据中台要解决什么问题

### 1.指标口径不一致

通常表现在3个方面：业务口径不一致、计算逻辑不一致、数据来源不一致。

- 业务口径不一致：业务口径不一致的指标，应该要有不同的标识去区分，比如上面提到的销售额这一指标，明明口径是不一致的，但却没有区分，容易让业务误解。
- 计算逻辑不一致：业务口径的描述往往是一段话，但对于一些计算逻辑比较复杂的指标，一段话通常是

描述不清楚的，如果碰巧两个相同业务口径的指标是不同的数据研发实现的，极有可能会出现计算逻辑不一致的情况。

- 数据来源不一致：对于部分指标，有多个数据源可供选择，如果数据源正好有些细微差异不被发现时，即使加工逻辑一样，也有可能结果不一致。另外，实时数据和离线数据也会有一定差异。

因此，要实现一致性，就要确保对同一个指标，只有一个业务口径，只加工一次，且数据来源必须一致。

## 2. 烟囱式建设数据平台，大量源被浪费，响应速度慢

主要在于烟囱式的开发模式，使得数据复用性低，导致大量重复逻辑代码的研发，影响需求响应速度。

比如，两个指标都需要对同一份原始数据进行清洗，原则上来说，只用一个任务对原始数据做清洗，产出一张明细表，另一个指标开发时，便可直接引用已经清洗好的明细表，这样便可节省一个清洗逻辑的研发工作量。但现实往往是对同一份原始数据做了两次清洗。因此，要解决需求响应速度慢的问题，就要提升数据的复用性，确保相同数据只加工一次，实现数据的共享。

## 3. 取数效率低

主要表现在两个方面，一方面是找不到数据，另一方面是取不到数据。要解决找不到数据的问题，就要构建企业数据资产目录，让数据使用者快速找到并理解数据。取不到数据的主要是非技术人员不会写SQL去提取数据，所以可以为其提供自助取数工具，使其简单快速的获取数据。

## 4. 数据质量低

面对业务已经沉淀的大量数据，逐步形成了企业的数据资产。而这些数据资产如何成为可持续使用的，为企业带来价值的数据，需要数据治理进行提升数据质量，比如设计数据质量校验的规则和使用流程，设计数据管控权限，数据如何安全输出及共享的设计等，如何在整体上发挥出数据的协同效应，为业务提供更高价值的数据服务链路，数据中台可以将这些数据能力整合到一起，对业务端提供稳定的持续的服务能力。

# 五. 什么企业适合做数据中台

数据中台的构建需要大量人力物力的投入，所以数据中台的建设一定要结合企业的现状，按需选择，不可盲目跟风。因此，企业在选择是否构建数据中台时，可以从以下几个方面思考：

**首先，看企业是否有一定的数据基础**，是否实现了业务数据化的过程，有了一定的数据沉淀，数据中台，顾名思义，数据是基础；

**其次，企业是否存在业务数据孤岛**，是否有需要整合各个业务系统的数据，进行关联分析的需求，如果有，需要通过构建数据中台，打通数据孤岛，整合各业务系统数据，满足关联分析的需求。比如某零售企业，在业务发展初期，商品、销售、供应链等都是独立的数据仓库，后期要构建智能补货系统，需要打通多个业务系统的数据，因此选择建设数据中台；

最后，在日常的数据使用过程中是否遇到指标口径不一致、需求响应速度慢、数据质量差、数据成本高等痛点，如果满足前两个条件，且在数据应用中存在以上所述的一些痛点，那建议你可以考虑将数据中台项目提上日程了。

# 六. 数据中台怎么建设

## 01 入手点

应从面向“业务价值”入手，简单来讲就是，面向应用更有目标性，能更早地发挥数据的价值，让企业客户的数字化转型路径不再是一个漫长的周期建设，而是一个逐步演进的过程。换一个更好的理解方式，其实是面向企业客户实际需求，以及业务价值构建数据中台。

首先，上数据中台的最好是业务发展或变化快速的部门，因为这些业务上中台，一是ROI容易成正比，二也能充分发挥数据的价值，容易得到各方认可；

其次，一开始不一定就得从统一数据口径入手，是不是可以先容忍数据层面一定程度的混乱，验证价值。当业务发展起来后，再去治理它，这很大程度上符合敏捷的理念，也符合很多企业的实际情况。

然后，针对业务价值或实际存在的问题提供服务，务实而非务虚。比如，①先上专家或架构师，进行项目诊断；②用产品和解决方案，走通关键路径；③当核心业务问题被解决后，也有一些事情是需要客户自己来完成，这时也能够针对性提供一些咨询服务。

## 02 匹配企业数字化进程

建设数据中台要遵循企业数字化进程各阶段的要求，因此，企业数字化发展可以分为**数据汇集、融合、开放、智能化**处理几个阶段。

第一阶段，对于本身已经覆盖较多信息系统的企业，需要考虑把有关数据汇聚到一起。而对于信息化程度相对偏低的企业，则要实现企业业务的在线化；

第二阶段，需要企业评估其自身数据是否已经实现了有机地融合。所谓的“融合”指的是企业通过一种标准把各个系统产生的数据进行有效的资产化。也就是说，这个阶段企业需要完成数据治理和归集工作；

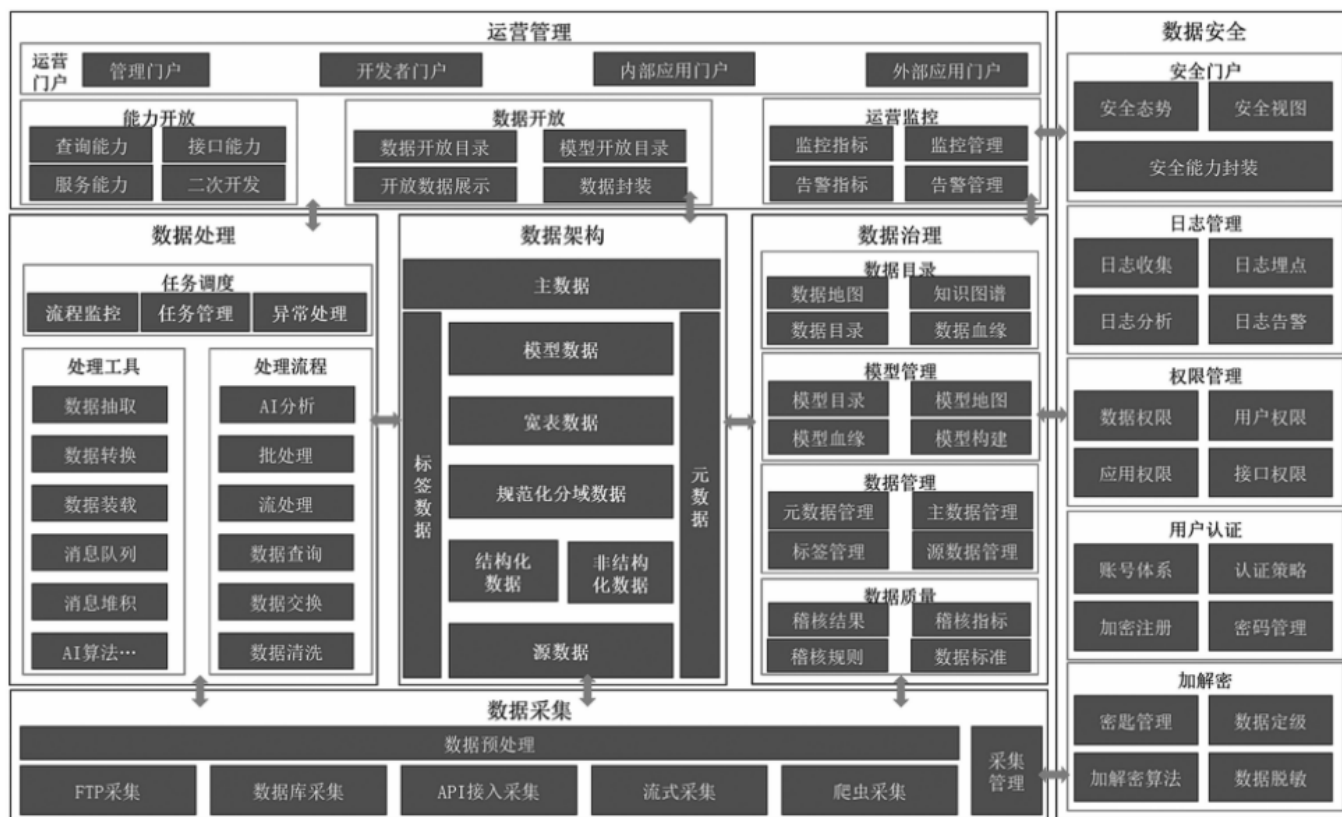
第三阶段，涉及数据的开放，即企业需要有专门的部门把归集以后的数据开放给内部各个部门，让各部门了解企业的数据资产情况，从而更好地实现企业基于数据的服务提升与创新。有条件的企业再把数据开放给生态链上下游的企业，实现服务创新、协作方式的重构，从而形成更大范围的协同；

第四阶段，指的是利用数据进行智能化处理。众所周知，企业通过机器学习等人工智能的方式进行数据处理，可以创造出十分广阔的增值空间，就像寻找矿产资源一样，通过数据智能的方式，企业可以从前所未有的角度挖掘出全新的数据价值。

以上的数字化进程对于计划实施数字化战略的企业而言，是相对比较适合的一个过程。同时，由于各企业的实际情况不同，各自的战略也会有所差别。大型企业建设中台主要需要考虑转体系问题，即企业应从整个组织、商业模式、战略协同方面，开展全面的改造，即三个全：全在线、全链接、全协同。而发展中国企业则需要先考虑“工具化”问题，即企业可以借助数据平台、工具，首先实现业务的在线化，然后再考虑基于数据的服务提升。

## 03 数据中台架构

从数据处理与数据治理两个维度出发，可以设计一个解耦的数据中台体系架构。该数据中台体系架构具有一定的柔性，可按照企业应用需求进行组合，或者对单个模块进行扩充，能满足大多数企业数据中台建设的需求。



数据中台的通用体系架构如图 所示。该中台体系架构以减少功能冗余和提高功能复用为原则，把数据中台解耦为 6 个可以分别独立建设、演进的功能子系统。

数据结构与数据处理子系统是数据中台体系架构的核心，数据治理是提升数据价值的重要手段。该数据中台体系架构的通用性表现在以下几点：

(1) 该数据中台体系架构综合考虑了数据中台的各种要素，参考这个架构进行建设可以有效提升数据资产价值，提供数据及服务的共享。

(2) 参考这个数据中台体系架构，企业可以一次规划、分步实施。首先建设处理子系统及数据存储子系统，然后根据业务发展需求，逐步补充数据采集、数据安全及数据治理子系统。

(3) 该数据中台由 6 个解耦的子系统组成。企业在立项建设时可以灵活组合，每个子系统单独招标建设，也可以把多个子系统合并招标建设。数据中台通用体系架构包含**数据采集框架、数据存储框架、数据处理框架、数据治理框架、数据安全框架及数据运营框架**等 6 大部分。

## 1) 数据采集框架

数据中台的采集框架应对纳入数据中台的各种源数据进行统一采集管理。数据采集框架中应提供多种数据采集方式，如**文件传输协议采集、数据库采集、接口应用程序接入采集、流式采集及网络爬虫采集**。

同时采集框架应按照数据采集规范对源数据进行预处理，从而去除明显不需要的数据及多余数据，并对采集过程进行管理。虽然数据中台的体系架构没有统一模板，但各企业数据采集框架基本一致。

## 2) 数据存储框架

数据中台的核心是数据，数据通过采集系统获取，然后数据经过处理框架加工，并接受数据治理框架的管理，同时也要接受数据安全框架的管理，最后开放的价值数据将通过数据运营框架对外提供数据服务。



数据中台的数据架构应该独立规划，并采用合理的技术架构对不同类型的数据进行存储。数据存储框架中，无论数据采用对象存储、块存储还是数据库存储技术，各种中台数据可按照上图所示分类管理。

源数据主要由采集框架进行管理，数据治理框架按照数据特征把数据简单分为结构化和非结构化数据两大类，而规范化分域数据则是数据治理框架对全量数据的规范化分域整理。宽表数据是数据关联的结果，利用宽表数据可以对人、事、地、物、组等对象进行完整的数据画像，同时宽表数据也可以作为上层模型数据的中间层数据。

元数据和标签数据都是对数据的描述，其中元数据用来对数据的客观属性进行表示，标签数据更倾向于管理者对数据的主观表述及等级划分，比如质量等级标签、安全标签、属性标签等。主数据需要在各系统间频繁更新、交换，且需要独立的存储空间进行维护管理。

### 3) 数据处理框架

数据处理是每个数据应用的基本环节之一，经典的数据抽取、转换和加载（ETL）处理流程在数据采集预处理、数据整合、数据建模等多个地方均要使用。单独建设数据处理框架有利于数据处理工具组件的集中开发与与管理，也有利于数据中台数据处理任务的协调与调度。

数据处理框架专门负责数据处理相关的任务，包括**批处理、流处理、人工智能分析、数据清洗、数据交换及查询**，此外数据处理的相关工具组件可在处理框架中配置。任务调度模块在数据处理框架中处于居中指挥的作用，并对运行的数据处理任务进行监控及异常处理等操作。

### 4) 数据治理框架

广义的数据治理不仅包含提升数据价值的内容，如数据管理、数据目录、数据质量等，也包含数据安全管理及数据共享服务。

数据安全管理与数据价值提升是一个矛盾体，如果由一个厂商或开发团队进行数据安全管理及数据价值提升相关软件的开发，则开发者的操作难免有所偏向，而且矛盾不容易公开，少了冲突也就少了优质的解决方案。

另外，数据共享与数据治理的其他内容也存在相同的问题。因此，本文建议数据中台的数据治理框架中不包含数据安全与共享的相关内容。

数据治理框架包含**数据资产目录、数据管理、模型管理和数据质量** 4 个模块：

（1）数据地图、数据资产目录、知识图谱及数据血缘的主要作用是展示数据的属性及相互关系，因此都纳入数据目录模块。

（2）数据模型能提高数据中台对外部应用需求的反应能力，固化的中间模型数据需要专门管理。模型管理包括模型目录、模型血缘及模型地图等。

（3）数据管理又可以细分为元数据管理、主数据管理、标签数据管理及源数据管理。

（4）数据质量管理模块按照制定的数据标准及数据稽核规则对数据中台的数据进行质量管理。

### 5) 数据安全框架

数据已经成为数据资产，数据安全框架是数据中台必不可少的组成部分。数据安全叠加在数据中台其他功能框架之上，数据采集、处理、交换、共享等每个环节均必须实施安全控制策略。安全框架可以分为日志管理、用户认证、权限管理及加解密等几个功能模块。

此外，安全全门户也可以对外提供安全能力封装，展示数据中台的安全态势及安全视图。

## 6) 数据运营框架

数据中台的核心功能是综合众多数据应用的数据处理及数据治理功能，集中建设、集中管理、减少冗余、增加复用。数据中台的最终目的还是为其他应用或开发者提供数据服务，而对外数据服务功能将直接面向不确定的外部对象。

因此单独建设数据运营，一方面有利于针对外部用户提供针对性功能；另一方面，数据运营模块作为用户与数据中台核心数据服务之间的中间层，可以有效隔离外部用户直接控制、接触核心数据及应用，可保护数据中台的安全性及内部功能的稳定性。

综合以上因素，数据运营应配置运营门户、能力开放、数据开放及运营监控等功能：

**(1) 运营门户：**对数据中台管理者提供管理门户，对开发者提供开发者门户。对内部应用提供内部应用门户，对外部应用提供外部应用门户。运营门户针对不同的用户提供不同的通道并开放不同的数据中台能力。

**(2) 能力开放：**把数据中台的数据处理能力、数据分析能力等经过适当的封装后对用户提供服务，可以是微服务，也可以是 API 接口，或者直接提供二次开发能力。

**(3) 数据开放：**通过数据目录，数据/模型展示（可视化、数据视图等）为其他数据应用系统提供数据服务。

**(4) 运营监控：**对数据中台的总体运营情况进行监控管理，包括硬件环境、软件环境，并且确定监控指标，按需求提供运营日报，处理告警信息。

## 七.数据中台发展趋势

### 1、标准化与市场下沉

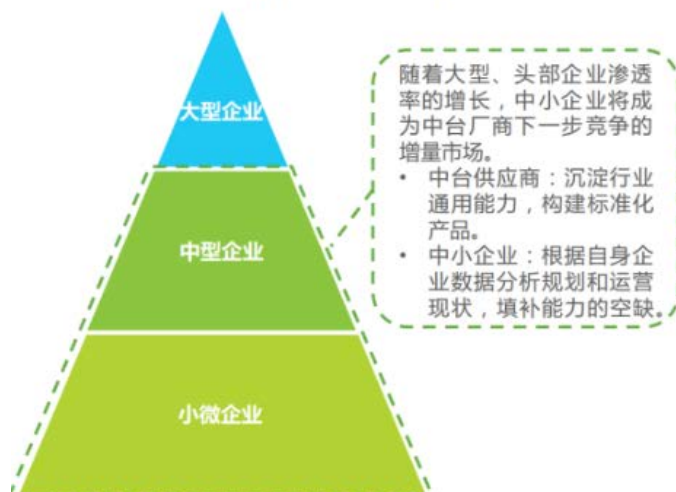
数据中台的核心在于共享和沉淀能力，随着数据中台在行业头部及领先企业逐渐落地，供应商经历了各类业务场景能力沉淀的过程。

在深度上，数据中台厂商承载细分行业的各类定制化业务，不断沉淀业务能力。

在广度上，随着不同业务场景的持续输入，数据中台厂商产品的能力越来越丰富，覆盖的领域也越来越广泛。

完善数据中台的深度和广度，提炼和整合数据中台的服务，尤其是对于对数据中台能力要求相对简单的中小企业，为客户提供标准化的整体解决方案将成为数据中台服务商的产品方向。

## 数据中台厂商发展下沉趋势



## 数据中台与产业链上下游企业关系



## 2、精细化

首先，数据中台所提供的底层技术支撑能力，需要供应商在软件架构、云技术、容器编排、DevOps等多方面有充足的技术储备，还需要具备资本和技术实力的双重积累。

纵观中国数据中台行业，虽然界限并不明晰，但是大致形成了以阿里、腾讯等技术雄厚的头部企业侧重提供底层架构技术，其他中小供应商侧重提供行业化服务和产品的竞争格局。

其次，没有一家供应商可以覆盖企业庞大的、所有的需求，尤其是多组织、多板块、跨业务的大型企业，所以在一个领域内已经完成实践和形成规模的供应商会优先深耕本领域，提供更加细分的场景切入口。

最后，企业也会根据业务需求面向不同领域的数据中台产品进行选择，不会局限于一家中台服务商。随着创业公司不断成长，细小赛道逐渐被填充，愈加激烈的市场竞争会使差异化成为供应商采取的产品战略。

## 数据中台主要细分领域



## 数据中台细分领域发展促进因素



## 3、SAAS化

从内部来看，数据中台不断沉淀跨行业、跨企业复用的组件、模块，存在朝SaaS和本地部署混合模式发展的趋势。从外部来看，随着云计算的普及，部分系统SaaS化趋势较强。因此，作为前台和后台的连接，数

据中台与SaaS应用融合对接的 实践越来越多，市场将逐渐形成一套成熟的中台+SaaS系统融合闭环方案。

敏捷开发、快速迭代以适应业务需求是数据中台的基本能力。随着数据中台市场渗透率的提高，应对小量应用调整的场景，低代码需求在近期兴起。允许通过零代码或少量代码就可以快速创建应用，对企业运维团队的要求降低，将充分提升数据 中台的应用性。



4、智能化

海量数据与多样的业务场景导致数据中台数据量大增，积累了丰富的数据指标，未来数据中台将会应用智能技术提供通用化智能服务，为业务决策提供直接辅助场景，比如商品销量预测，千人千面推荐算法、营销活动预测等。同时，通过智能技术算法可以为前端员工降低数据使用的门槛，提高整体工作效率和生产效率。

【END】