# 第8章

# 数据智能平台构建策略

大数据是数字化的产物，随着业务成熟度的逐渐提高，面对的需求逐渐多样化和个性化，对于创新的要求也越来越高，可以说智能数据是大数据发展的高级阶段，是大数据在应用创新落地方面的核心要求。本章将介绍大数据智能平台的建设内容，建设的核心过程包括系统、业务、平台三方面的设计思路、建设体系和实施方式。本章分为三部分，第一部分介绍数据业务的构建，第二部分介绍系统＋平台如何构成数据智能的体系，第三部分结合目前最新的数据中台的概念进行讲述。

## 8.1 数据业务的构建过程

通用的开展大数据业务的过程如图 8-1 所示。


图 8-1 大数据业务构建过程

首先是数据系统的建设，数据系统是基础。从确定要进行哪些方面的数据收集开始，需要把收集到的数据进行清洗、筛选、格式转换，然后存入系统，并且按照技术平台的要求投入人力、设备等进行大数据系统的搭建。其次是数据业务建模。有了系统，就可以基于这个系统来观察数据，可以由建模人员利用其专业知识基于机器学习来进行建模，在得到一个合适的模型之后，需要把此模型放到大数据系统中运行。一般来说，这个大数据系统需要有大数据工程师一起参与，将模型转换成适合在平台上运行的代码，后面逐渐地会出现很多高效率

的工具来帮助这种代码化的转换。最后是数据业务开展，需要把数据价值体现到业务上去，也就是数据业务的发展，通过分析人员对数据进行再整理、可视化呈现、洞察后指导业务开展。而如果从中可以抽象出新的产品，就可以通过产品设计来形成创新，创造出新的商业价值。

### 8.1.1　数据系统建设

为了把数据系统建设讲清楚，特别是把其中的要点、难点等清晰地呈现出来，下面采用一个现实中的基础建设的例子来说明。

假设目前需要在一个靠近大海的地方建设一个新型设备工厂，这个设备可以用于日常生活中，会极大地提高我们的生活水平，但是目前市场的前景不是特别明朗，而建造这个设备工厂所需要的原材料很大一部分又需要从各分散的城市或者城镇中运送过来。

作为工厂进行生产制造的基础，我们需要建造公路来连接原料产地和工厂，也需要建造厂房来进行生产，也就是需要基础设施的建设，那么对于大数据技术来说，大数据系统建设就属于基础建设要求。

依据对于市场的认识以及资源（资金、能力等）的准备情况，建设基础设施（以构造公路作为主要的工作为例）必须明确以下几点：

- 造路的主要目的是什么？
- 从哪里到哪里、中间有多少出入口？
- 什么时间满足多少交通流量（阶段、造多宽的路、车辆类型、可以运载什么货物、允许最大数量等）？
- 目前拥有的资源是什么（预算、团队、时间等）？
- 阶段性的规划是什么（资源、目标、实施）？

这时最主要的一点就是要清楚造路的主要目的，也就是建设这个系统的近期、远期目标是什么？可以根据 7.5 节中的“成熟度模型”进行规划。这一目的也是图 8-1 中最上面的部分决定的。在此目标指导下，需要盘点有哪些城市、城镇需要接入这个公路系统。这时难点就在于梳理以下几个方面：

- 哪些城市需要接入（也就是需要哪些原料、生产出来的设备会运往哪里）？
- 这些城市到达各入口的支路是否建设好？
- 建设这些支路对于原有系统的影响多大？
- 如果影响比较大的话，如何解决？
- 原料是否还需要再加工？
- 原料的量是多少？

这些城镇就好比公司中不同的业务系统，对应到大数据系统，下面就是需要解决的问题：

- 是否确定了数据源头对应的业务系统？
- 这些系统通过何种方式来准备数据？
- 数据如何被接入大数据系统？
- 源数据是否已经被收集？
- 数据格式是否已标准化？

- 数据量是多少？

把城市通往工厂的路造好后，并不是就一劳永逸了，后续依然需要根据需求不断建造、维护、升级。同时还需建造厂房、购置生产设备、建立流水线、建造仓库用于存放原料和生产出来的设备等。

对应到大数据系统建设方面，包括以下几项内容：

- 数据收集系统：确定数据源、数据格式、数据传输方法、数据清洗工具等。
- 搭建存储集群：确定存储规模、服务器配置和数量、网络规划及建设、安装和调试集群、确定存储方式等。
- 搭建计算集群：确定计算方式、计算规模、服务器配置和数量、网络规划及建设、安装和调试集群、任务调度机制等。
- 数据安全策略设计(可以按阶段进行)。

### 8.1.2 数据业务建模

在把厂房、流水线等初步建设完成后，就可以把所需要的材料经过多种方式运送到工厂，接下来就需要有一些专业的工程师进行以下活动。

- 为了保证后续生产的效率，需要对原料进行分门别类，确定存放地点和存放顺序，必要时还需要进行一定的搭配。
- 从这些材料中挑选出一些进行化验，确定其成色和质量，最后确定哪些可以用，哪些不可以用。
- 进行加工工艺的设计，哪些材料什么时候通过什么方式进入生产线，哪些零件先生产出来，哪些零件后生产出来，如何装配。
- 对生产出来的设备确定调试和验证方法，确定其在质量要求范围之内。

这个工作对应到大数据技术中就是数据建模。数据建模就是建立数据存放模型并处理，把各数据源的各种数据根据一定的业务规则或者应用需求对数据重新进行规划、设计和整理。然后根据产品的要求，利用这些数据的样本进行模型的建立，确定输入的数据要求，送入处理流水线，一直到产生最终的结果。

这个阶段的难点和要点在于：

- 需要有具有行业专业技能的人才，这类人才首要的能力是具有行业相关的业务知识和洞察能力，掌握行业内常用的建模经验。
- 特征工程，确定哪些特征可以用于业务模型。由于数据在收集过程中，数据输送方由于各种原因，事先并不一定清楚或者预见会服务于何种业务，而在实际使用时需要进行再处理(标准化)以满足建模的需要。所以对于各种形式的数据，需要通过特征工程来进行特征筛选、特征组合、特征变换等，才能为后续的模型所使用。
- 为数据确定高效的存取模型。经过特征工程后的数据可以作为模型的输入进行建模，为了保证在生产环境中的模型的运行效率，需要确定数据的存取模型，还需要进行宽表、数据仓库的设计和构造，否则会导致资源的浪费。
- 模型架构的确定。采用流式处理还是批量处理，采用何种调度方式，需要多少运算资源，输出结果如何存放等，也是一个难点和要点。

下面讲述 AI 建模的方法论。建模过程中使用 AI(机器学习技术)作为内核能力，其过

程如图 8-2 所示。

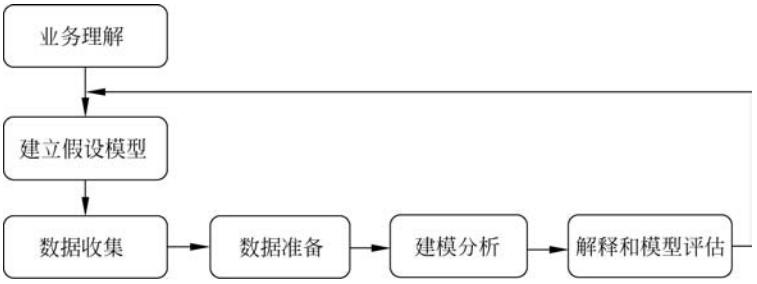


图 8-2　AI 建模的一般方法论

**1. 业务理解**

把业务问题理解透彻，理解项目目标和需求，将目标转换成问题定义。

难点：需要对业务领域有比较深入的理解，而且不仅仅是业务专家，还需要具备数据和技术感觉。

**2. 建立假设模型**

设计出达到目标的一个初步计划。根据直觉和知识提出合理假说，如类比相关性等。

难点：如何设计合理的目标函数，使其达到业务初始设计要求。

**3. 数据收集**

收集初步的数据，进行各种熟悉数据的活动，包括数据描述、数据探索和数据质量验证等。要有数据，而且的确需要足够多的数据。

难点：如何解决数据收集成本大的问题，或者说如何自动化收集数据。需要收集多少数据才够，学术界尚未有固定的理论指导，一般从成功案例中提炼经验公式。

**4. 数据准备**

需要首先弄清楚数据来源，然后进行探索性数据分析（Explore Data Analysis，EDA）去了解数据的大体情况，通过描述性统计方法提升数据质量，将最初的原始数据构造成最终适合建模工具处理的数据集，包括表、记录和属性的选择，数据转换（稀疏，异构）和数据清理（缺失、矛盾）等。

难点：对于优质数据的判断标准等。

**5. 建模分析**

选择和应用各种建模技术，并对其参数进行优化。一般情况下，为了让模型更好地达到效果，在偏差和方差方面得到最优结果，常常把数据集分为两部分，一部分用于开发训练（训练集、验证集），一部分用于预测（测试集）。

难点：算法和参数如何选择，目前选择是根据类比的方法，寻找与待解决工程相似的已成功的工程，并使用相似的方法，但工程相似没有统一标准。对于参数的选择，目前常用方法还是尽可能多的实验，选择测试结果最好的参数。

**6. 解释和模型评估**

对模型进行较为彻底的评价，并检查构建模型的每个步骤，确认其是否真正实现了预定

的目的。

难点：目前还没有对于效果不好的原因定位的方法，只能具体案例具体分析。

### 8.1.3 数据业务开展

设备生产出来以后，就涉及设备投放到市场，卖给消费者，做好服务。然后根据市场反馈，对产品进行改良、升级（创新），同时还需要让公司的各个部门能及时获得产品的表现和市场要求。

从服务于客户和市场的角度出发，数据产品或者数据业务本身也是数据的来源，这些数据依然需要通过大数据平台来对产品质量、用户互动产生的反馈信息进行收集处理，同时需要把信息及时展示和传达给各个部门。

客户的要求是多种多样的，无论是内部客户还是外部客户，所以根据数据客户的需求不同，数据业务开展的形式也不同：

- 老板们（或公司管理层）时间宝贵，注重宏观，一般只看重要指标，并且要求图文并茂、简单易懂。这就好比餐馆所有菜品都是固定的，但是菜品得色香味俱全，上菜速度得快。所以大厨们得事先把数据加工成仪表盘、可视化大屏等让人对关键指标一目了然、卖相高大上的数据应用，并且采用各种技术手段保证数据应用的性能。
- 各部门主管每天都要面对各种日常工作和突发情况，所以他们对数据的要求是既要能满足日常管理需要，也要能有额外的手段来应对突发情况，而且这些手段速度不能慢，毕竟服务是 7×24 小时不间断运行，所以需要将数据加工成多维分析、自助分析一类的数据应用，根据经验和主管们的业务需求，将有可能用到的东西全部提供出来，可以根据需要随意使用。
- 一般客户（或者员工们）也有数据需求，但通常需求简单，难点在于人多、需求量大，所以将数据加工成报表这种类似于快餐的数据应用是最好的方式。

这个阶段的难点和要点在于：

- 如何形成数据业务本身开展过程中的数据处理的闭环。
- 针对不同的客户，形成不同的数据维护和可视化等工具。
- 满足各种数据需求基础上的数据创新。
- 数据分析师、数据科学家等角色的物色和参与。

## 8.2 数据智能体系要求

本节讲述如何从技术体系上进行建设。首先就建设思路、原则和目标进行讲解，然后搭建基础平台来进行系统治理和系统保证，业务目的是进行数据挖掘计算，从数据中挖掘知识来支撑业务决策。在执行过程中需要保证数据安全及隐私，最终通过可视化系统以用户容易理解的方式展示出来。这是一个自底向上的完整流程。

### 8.2.1 建设思路、原则和目标

经过近十年的发展，越来越印证了《大数据时代》一书中总结的以下几个核心观点。

- 改变操作方式，使用收集到的所有数据，而不是样本。
- 不把精确性作为重心。
- 接受混乱和错误的存在。
- 侧重于分析相关关系，而不是预测背后的原因。
- 数据的选择价值意味着无限可能。
- 数字时代要求我们对待数据有别于传统资产。
- 数据的创新意味着很大的不确定性。

总而言之，需要关注的核心点是如何面对数据创新的不确定性。

数据智能的定义中明确把数据定义成生产资料，然而这个生产资料和其他的生产资料有明显的不同，特别是以下几方面：

（1）数据不可知：用户不知道大数据平台中有哪些数据，也不知道这些数据和业务的关系是什么，虽然意识到了大数据的重要性，但平台中有没有能解决自己所面临业务问题的关键数据？该到哪里寻找这些数据？这些都不可知。

（2）数据不可控：数据不可控是从传统数据平台开始就一直存在的问题，在大数据时代表现得更为明显。没有统一的数据标准导致数据难以集成和统一，没有质量控制导致海量数据因质量过低而难以被利用。而且没有能有效管理整个大数据平台的管理流程。

（3）数据不可取：用户即使知道自己的业务所需要的是哪些数据，也不能便捷地拿到数据，相反，获取数据需要很长的开发过程，导致业务分析的需求难以被快速满足，而在大数据时代，业务追求的是针对某个业务问题的快速分析，这样漫长的需求响应时间难以满足业务需求。

（4）数据不可联：大数据时代，企业拥有着海量数据，但企业数据知识之间的关联还比较弱，没有把数据和知识体系关联起来，企业员工难以做到数据与知识之间的快速转换，不能对数据进行自主地探索和挖掘，数据的深层价值难以体现。

笔者对公司内部数据业务开展过程中的问题进行收集和汇总后，发现存在以下五大难点。

（1）对业务需求响应速度慢。

（2）数据质量问题频发。

（3）数据使用难以及获取数据慢。

（4）开发效能低，试错成本高。

（5）数据能力重复建设。

笔者认为数据智能体系建设的总体目标如下。

（1）敏捷地支撑业务部门的业务创新需求，打造快速服务商业需求的服务能力。

（2）把不同域的数据实时打通，体现数据的最大价值。

（3）把数据作为资产进行管理。

（4）直接的价值体现是成本节约、效率提升和质量提升。

数据智能体系的建设思路和原则如下。

（1）主要面向内部客户，特别是研发人员及建模人员，以提高业务开发效率为目标。

（2）做好元数据、血缘关系管理，提高数据治理程度，保证数据的质量和安全。

（3）提炼公共服务能力，复用程度高的能力优先建设。

（4）数据能力原则上由相应领域业务熟悉、技术积累强的团队一起参与建设。

(5) 能力建设需要重点考虑稳定、易运维、可运营、可审计。

图 8-3 所示为数据智能技术体系构成，至少需要包含基础平台、融合平台、治理系统、质量保证、安全计算、分析挖掘、数据可视化这几个方面。


图 8-3 数据智能技术体系构成

### 8.2.2 基础平台

目前，基础平台主要涉及以 Hadoop 为主的大数据平台的开发和建设工作，此部分将在第 9 章专门进行讲解。

### 8.2.3 融合平台

企业内部有不同类型的数据，同时也会从企业外部获得数据，这些数据会存在格式、定义、语意、编码等方式的不同，如何有效整理、融合如此多样且繁杂的数据对于数据智能平台非常重要。数据融合的相关技术在整体上需要解决以下关键问题。

首先，在机器从数据中获取智能之前，机器能够正确地读懂各种各样的数据。对于机器友好的数据是类似关系数据库的结构化数据。然而，现实世界里存在着大量的非结构化数据，例如自然语言的文本；还有介于两者之间的半结构化数据，如电子表格。目前，机器还很难理解这些非结构化的数据，需要将数据处理成对机器友好的结构化数据，机器才能发挥其特长，从数据中获取智能。非结构化数据，尤其是半结构化数据向结构化数据的转化，是实现数据智能不可或缺的先决任务。

其次，数据并不是孤立的，数据智能需要充分利用数据之间存在的关联，把其他数据源或数据集所包含的信息进行传递并整合，为数据分析任务提供更丰富的信息和更多的视角。

最后，数据并不是完美的，提前检测并修复数据中存在的缺失或错误，是保障数据智能得出正确结论的重要环节。

在功能范畴上，也可以把这部分归入数据治理部分。

### 8.2.4 治理系统

在数据智能中把数据作为核心资产和生产资料来看待，那么对于数据的治理即是重中之重。数据治理主要解决以下几个问题：

- 有什么数据(资产)。
- 如何确定数据的质量指标并保证数据的质量。
- 如何让业务使用方快速获取数据服务。
- 数据资产所有者如何向数据使用者进行科学的授权及监督。

再细化一下，应该从以下几个方面着手，给其余的系统提供支持。

- 治理结构方面，管理企业拥有的数据目录、数据类型，对组织结构设置相应的权限。
- 治理策略方面，能确定分类数据的敏感度水平，定义数据质量和数据标准要求，能设定对敏感数据进行脱敏(去标识化)的策略，明确定义数据共享等过程。

- 隐私和安全方面，能让用户控制对于隐私数据的授权，提供和管理对数据的访问，防止未经授权的访问，提供审计手段。
- 数据质量保证方面，通过构造数据地图、数据血缘图谱，在每一个数据结果上都可以回溯数据产生的细节，并准确定位问题所在。这在数据量、数据种类变得繁多时有绝对的必要性。

综上所述，数据治理是数据智能的基础，治理的好坏决定了数据智能的发展高度。

### 8.2.5　质量保证

数据的质量决定了数据产品和服务的质量，所以数据质量保证系统是数据治理系统基础上另一个重要的环节。数据质量保证可以从如图 8-4 所示的几个方面进行。

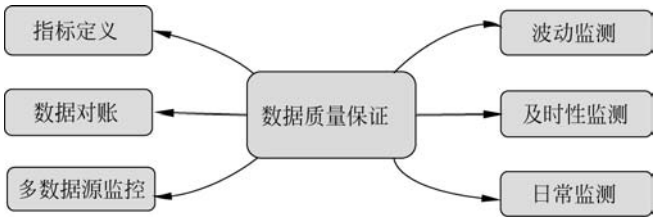


图 8-4　数据质量保证系统

首先需要明确指标，包括数据本身的指标和监控的指标。也就是定义尺子，只有尺子合理了，衡量结果才是稳定的，才能确定数据是否一致，是否出现异常波动，及时性是否达到要求。监控结果以可视化的方式呈现，做到信息全面展示。

其次要结合数据治理系统中的数据血缘关系、上下游关系，在监控到问题后，能及时进行问题定位，并快速采取措施纠正质量问题。

### 8.2.6　安全计算

除了数据治理外，还需要考虑如何让数据发挥更大的价值，如何能找到合适的合作者来联合创造价值，但是数据不同于别的资产，其具有可复制、难确权的特点，这就涉及目前行业内比较关注的安全计算技术。这方面的内容涉及多种加密技术、多方安全计算、同态加密、可信计算环境、数据隐私保护技术以及区块链技术。

### 8.2.7　分析挖掘

数据分析是数据智能中最核心的部分，大致可以分为描述性分析、诊断性分析、预测性分析、指导性分析四个类别，每个类别基于数据解决不同的问题，难度越大，所能带来的价值越高，所使用的技术也越复杂。

关于数据分析和挖掘系统在第 10 章专门进行讲解。

### 8.2.8　数据可视化

数据可视化本质上是为了感知和沟通数据而存在的，涉及不同的领域，如人机交互、图

形设计、心理学等。在当前大数据盛行的时代，数据可视化逐渐崭露头角，扮演着越来越重要的角色。

可视化技术已成为数据智能系统不可或缺的部分，这些技术通常会集成在一个图形界面上，展示一个或多个可视化视图。用户直接在这些视图上进行搜索、挑选、过滤等交互操作，对数据进行探索和分析。可视化工具逐渐趋于简单化、大众化，使一些高阶的分析变得更加简单。一些高级的可视化设计，如 Word Cloud、Treemap、Parallel Coordinates、Flowmap、ThemeRiver 等，已逐步成为主流。

在决策过程中，可视化也发挥着重要的作用，它能将信息展示得更准确、更丰富、更容易理解，从而极大地提高了人与人之间的沟通效率。可视化叙事（Visual Storytelling）研究如何将可视化用于信息的展示和交流。当今主流的数据分析平台，如 Power BI、Tableau、Qlik 等，都提供了可视化叙事的模式。可视化叙事的研究目前还处在初级阶段，人们还在探索它的各个方面，包括修饰形式、叙事方式、交互手段、上下文、记忆性等。如何评估一个可视化叙事也有待进一步研究。

随着数据业务的开展，数据资产不断丰富，需要我们的技术体系能更好为业务服务，以便快速响应，灵活组合。因此，一个有效的方法就是实施大数据中台策略。

## 8.3 数据中台策略

近几年数据中台的概念变得非常热，就本质而言，数据中台是大数据智能化的一种实施策略。

在数据中台之前，一直使用数据仓库（Data Warehouse）、数据湖（Data Lake）的概念，下面讲述这三者的提出背景和差别。

### 8.3.1 数据仓库和数据湖

许多公司已经选择“数据湖（Data Lake）”作为把所有数据收集起来的手段。数据湖的概念是于 2011 年提出来的，数据湖示意图如图 8-5 所示。或许是出于对数据没有保存而丢失的担忧，一些大数据厂商在 Hadoop 为基础的技术栈上，把一个组织中产生的原始数据存储在一个单一的系统中，一般大家使用开源的 Hadoop 来构建数据湖，不过数据湖的概念比 Hadoop 更广泛。那么数据湖与数据仓库或者数据集市的区别在哪里呢？

数据湖存储数据源提供者提供的原始数据，没有对数据的形式进行任何假设，每个数据源可以使用其选择的任何形式，最终数据的消费者根据自己的目的来使用数据。相对于数据仓库，这是一个非常重要的步骤，也是数据仓库没有走得更远的原因，因为数据仓库首先需要考虑数据方案（schema），如图 8-6 所示。

数据仓库倾向于为所有的分析需求设计一个总体的方案，但是实际上即使是一个非常小的组织，想要通过一个统一的数据模型来涵盖一切，也是不太实用的。另外一个数据仓库使用中的问题是数据质量，不同的分析需求对数据的构成有不同的质量要求和容忍度。数据仓库的这个特征，导致其漫长的开发周期，高昂的开发、维护成本，细节数据丢失等问题。

由于数据湖直观上更像一个数据质量差异很大的数据倾倒场，也因此产生了一个新的比较热的头衔：数据科学家，虽然这个头衔有点被滥用或者夸大，但是其中的许多人确实具

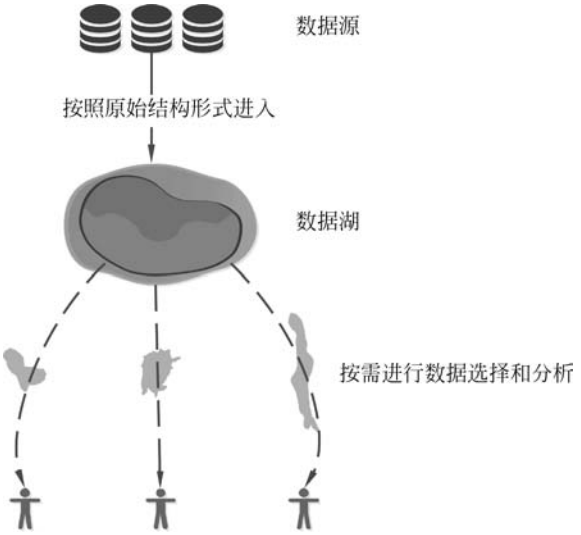


图 8-5　数据湖示意图

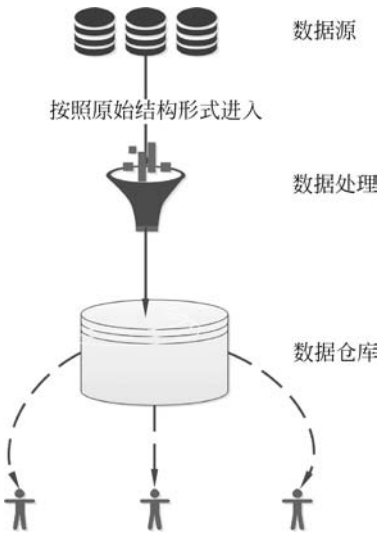


图 8-6　数据仓库示意图

有扎实的科学背景，并且掌握所有关于质量问题的知识，他们善于使用复杂的统计技术来找出数据质量问题，这为利用数据湖而不是以前一些不透明的数据清理机制来解决实际分析问题创造了条件。虽然数据仓库经常不仅仅只是数据清理，同时会将数据聚合到某种形式以便于被分析，但是数据科学家反对这种做法，因为聚合也就意味着丢掉很多数据。数据湖应该包含所有数据，因为不知道什么人通过哪些数据可以找到有价值的东西。

数据湖的这种原始数据的复杂性意味着用户可以通过某些方式来将数据转变成一个易于管理的结构，这样还可以减少数据的体量，更易于处理。数据湖不应该经常被直接被访问，因为数据是很原始的，需要很多技巧才能让其变得有意义。一般可以按照图 8-7 所示来处理，我们把它称为数据湖岸集市。

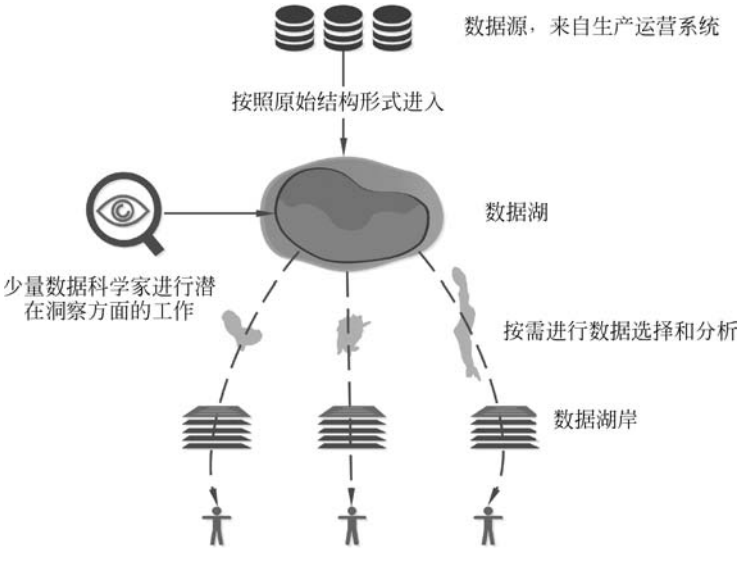


图 8-7　数据湖岸集市

把所有数据放入湖中的一个关键点是需要有一个清晰的治理。每个数据项应该有清晰的跟踪，以便知道从哪个系统中来以及数据什么时候被产生，等等，也就是元数据管理、数据血缘以及必要的数据安全。数据湖的一般处理流程如图 8-8 所示。

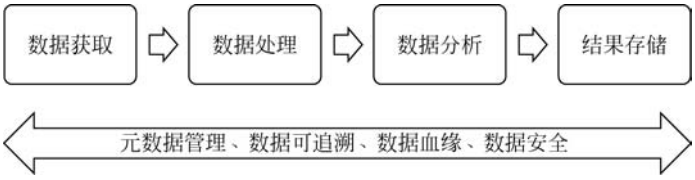


图 8-8　数据湖的一般处理流程

**1. 数据获取**

尽量获取最原始的数据，数据在获取过程中成为数据湖的一部分；数据可能以不同形式存在，也可能需要不同的机制来获取。

**2. 数据处理**

获取的数据需要进一步处理才能得到有用的信息，如进行画像、商品推荐、业务洞察力等，此时可能会用到机器学习技术。

**3. 数据分析**

数据进一步被分析，以便按需访问；数据分析需求受信息访问模式驱动。

**4. 结果存储**

数据分析结果需要存储在合适的数据存储系统中；数据湖中的数据存储系统的选择依赖具体的数据服务需求。

很多时候，数据湖被认为与数据仓库是等同的。实际上数据湖与数据仓库代表着企业想达成的不同目标，两者的关键区别如表 8-1 所示。

**表 8-1　数据湖与数据仓库的对比**

| 比较项 | 数　据　湖 | 数 据 仓 库 |
|---|---|---|
| 处理对象 | 能处理所有类型的数据，如结构化数据、非结构化数据、半结构化数据等，数据类型依赖于数据源系统的原始数据格式 | 只能对结构化数据进行处理，而且这些数据必须与数据仓库事先定义的模型吻合 |
| 数据用途 | 拥有足够强的计算能力，能处理和分析所有类型的数据，分析后的数据会被存储起来供用户使用，也就是 Schema-On-Read，在使用时才需要给予定义，因而提高了数据模型定义的灵活度 | 处理结构化数据，将它们或转化为多维数据，或转换为报表，以满足后续的高级报表及数据分析需求，也就是 Schema-On-Write，模型是在数据被写入之前就定义好的 |
| 应用场景 | 数据湖通常包含更多的相关的信息，这些信息有很大概率会被访问，并且能够为企业挖掘新的运营需求 | 数据仓库通常用于存储和维护长期数据，因此数据可以按需访问 |

### 8.3.2　数据中台

企业的前、中、后台如图 8-9 所示。

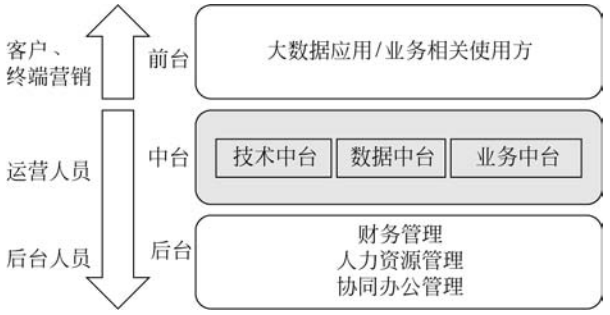


图 8-9　企业前、中、后台

**1. 前台**

前台也就是面向客户(customer)的系统，这里的客户可以是 toC 的，也可以是 toB 的。例如电商网站、微信店铺或者是面向渠道商的门户。

**2. 业务中台**

业务中台对后台的系统资源进行整合、封装，转化为前端可以使用的公共服务。

业务中台是前端应用所需服务的提供者，前端应用是服务的消费者，二者相辅相成，共同发展，业务能力不断沉淀到业务中台。这里要说明一点，业务中台不必拘泥于一种形式，它可以是一组无界面的公共接口服务、也可以是一个独立的系统或有界面的工具。

例如淘宝商品中心就是一个非常庞大的体系，既有商品搜索、排序等接口服务，也提供各种直接操作的工具，商品发布直接提供用户操作界面，商品的类目管理也有淘宝小二的操作界面。

**3. 数据中台**

数据中台从后台及业务中台对数据进行抽取，完成海量数据的存储、计算、产品化包装过程，构成企业的核心数据能力，为前台基于数据的订制化需求提供支撑。

**4. 技术中台**

技术中台一般指底层 PaaS 的能力，PaaS 层主要解决大型架构在分布式、可靠性、可用性、容错、监控以及运维层面上的通用需求。

技术中台对互联网公司尤其重要，也就是我们俗称的“三高”，在高可用、高性能、高并发方面，技术难度高、投入大，不可能为每个业务都做一套系统。

**5. 后台**

后台简单来说就是支撑公司业务开展的公共的职能模块，如财务管理、人力资源管理、协同办公管理等。

### 8.3.3 数据中台和数据仓库、数据湖的差别

数据中台与数据仓库、数据湖最大的区别就是数据中台更加贴近业务，不只提供分析功能，更重要的是为业务提供服务，与业务中台或者业务系统联系更加紧密。

数据和业务中台的建设更多的是从前台业务的角度进行提炼，提炼出可以复用的组件，形成业务中台，然后再进行数据的组织，结合采集的数据和业务开展中积累的数据，通过公共数据接口等形成数据中台，最后通过技术中台来落地。

下面举例来说明数据中台和数据仓库、数据湖的差别。以大家熟悉的“千人千面”案例来说，需要根据不同的客户展示不同的推荐内容，除了要整合业务系统产生的用户基本属性、订单、评价、加入购物车等行为数据外，还要通过埋点的方式实时获取用户的偏好浏览、搜索、分享商品等行为数据，经过数据中台一系列的数据加工处理后，最终以微服务的形式提供。

在业务系统每个需要给目标用户呈现商品的数据服务处，已不是简单地、一成不变地去商品库查询数据，而是调用数据中台提供的商品推荐接口，以此根据不同人的偏好、浏览历史、商品相似度等数据来为每个人推荐最感兴趣的商品。这种业务、数据紧密联动的场景在数据仓库时代是完全做不到的。

数据中台与数据仓库、数据湖另一个差别是企业是否把数据作为一种单独的财产进行组织和管理，这就涉及企业组织结构是否有相应的配置，而不是仅仅作为一种 IT 系统来对待。

# 第9章

# 大数据技术和平台

在企业的大数据业务开展中，技术平台虽然不是决定因素，但绝对是一个必要的基础。虽然最终大数据业务是否成功并不仅仅依赖技术，但是没有技术平台肯定存在大问题。通过从上到下贯彻大数据意识以及数据价值观念，采用一定的方法论，依靠必要的产品、技术平台、工具等，用制度流程等执行起来，才是一个完整的大数据价值实现体系。

本章主要就大数据技术平台进行阐述。考虑到本书的读者主要是高级技术人员或者公司管理人员，所以较基础的技术内容不会过多展开。

## 9.1 大数据基础技术系统组成

传统的操作系统组成如图 9-1 所示，图中的七部分组成了一个完整的操作系统，大数据技术由于其特性，基本无法通过单设备进行处理（有一些商用大型计算机或者超级计算机可以处理某些应用场景），因此需要通过成千上万的服务器以集群的方式来完成，单体的系统需要扩展到集群系统。从管理学来看，群体（集群）的管理是一门比较复杂的学科。

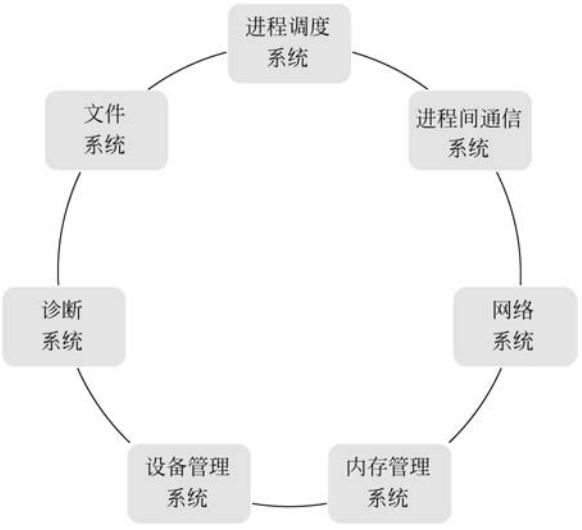


图 9-1 操作系统组成