

# 第 1 章

## 认识数据资产

在当前以数字化为主导，人工智能技术日益成熟的信息技术时代，大数据技术已成为企业和国家的核心竞争力。早在2015年，国务院便发布了《促进大数据发展行动纲要》，着重强调了要加强数字政府建设、加快推进全国一体化政务大数据体系建设。数据不仅对于国家的建设和发展来说非常重要，是很多企业或者组织决策的重要依据，更是驱动企业或者组织自身业务增长和创新的关键要素。企业或者组织通过对数据资产的管理，以及通过大数据技术对数据进行分析和挖掘，可以更加深入地了解市场需求、提升产品的服务以及运营效率，使得在激烈的市场竞争中不被淘汰。

### 1.1 数据资产的基本介绍

数据资产通常是指那些可以通过分析来揭示价值、支持企业决策制定、优化企业流程、预测行业的未来趋势或产生更大的经济价值的数据集。这些数据可能是由企业自身产生的，也可能是从外部获取的（如社交媒体、第三方数据提供商、网络爬虫等），而且这些数据的格式多样，可能是结构化数据、半结构化数据或者非结构化数据，如图1-1所示。

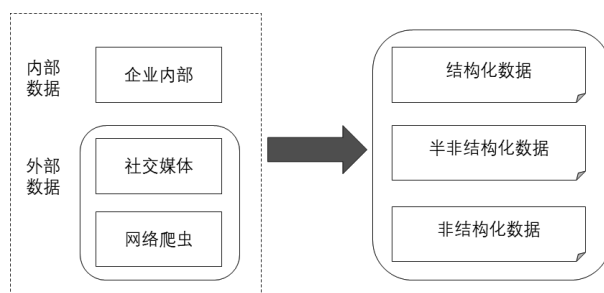


图 1-1

#### 1. 数据资产的关键特性

数据资产的关键特性是其可用性、可访问性、完整性、可靠性和安全性，通常这些特性共同决定了数据的质量和 value，如图1-2所示。

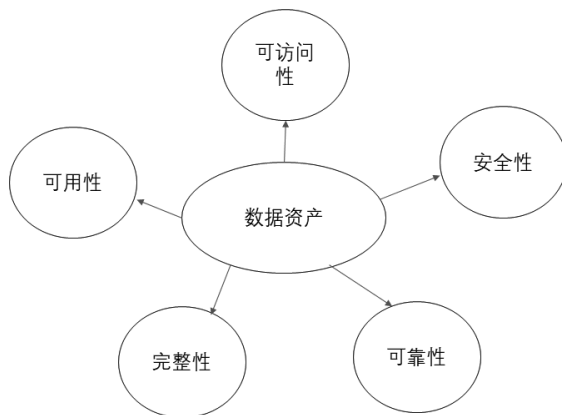


图 1-2

- 可用性：指的是数据资产需要能够被使用。如果无法被使用，那么数据资产就无法体现其核心价值，而数据资产的可用性需要依赖数据质量、数据监控等很多关键要素的支撑。
- 可访问性：指的是数据资产需要能够被数据的使用者访问。如果无法被访问，那么数据资产会显得没有任何价值，因为只有能被访问，才能挖掘出数据的更多价值。
- 安全性：指的是数据资产需要保障其数据的安全性，防止数据被泄露、丢失或者被黑客攻击篡改等。
- 可靠性：指的是数据资产一定是可靠的，否则无法用于企业的决策和判断。如果数据不可靠，那么通过数据做出来的决策肯定也不会可靠，从而会给企业带来巨大的损失。
- 完整性：指的是数据资产中的数据一定是完整的。如果数据不完整，那么获取到的信息也不会完整，不完整的数据是无法用于数据分析、数据决策的。

但是在现实生活中，数据资产的价值往往不会立即显现。相反，它们通常需要适当的管理和分析，才能转换为实用的价值或带来直接的经济回报。如图1-3所示，例如，一个顾客在京东的商品订单数据，在原始形态下可能是一系列购买商品的交易记录，但是当通过数据分析揭示出消费者的行为模式和偏好时，这些数据就转换为有助于推动销售和制定营销策略的宝贵数据资产。

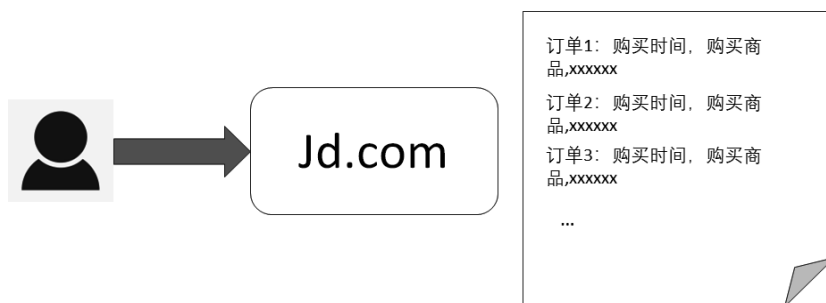


图 1-3

除用于内部决策支持外，数据资产还可以成为一种可以对外出售或交换的商品。随着数据市场的发展，越来越多的公司认识到通过共享或出售其数据资产可以获得额外的经济收入，或者与合作伙伴交换数据以获取共同价值。

随着技术的进步，尤其是大数据以及人工智能和机器学习的发展，数据资产的潜在价值正在急剧增加。通过数据挖掘以及机器学习的模型训练，可以进一步发掘数据的更高价值。

## 2. 数据资产的常见类型

以下是数据资产的几种常见类型。

- **结构化数据**: 这类数据通常存在于预定义的数据模型中，它们格式清晰、易于搜索和组织。结构化数据通常存储在关系数据库中，如SQL数据库，这类数据库支持复杂的查询、报告和分析。例如，客户信息、销售记录、库存数据和金融交易数据等，都可以以结构化的形式存储。通常它们以表格形式存在，每一列代表一个数据字段，每一行代表一个数据记录，如图1-4所示。

id	name	...	...										
1													
2													
...													
...													

图 1-4

- **非结构化数据**: 非结构化数据没有预定义的格式或组织，因此更难以处理和分析。这类数据包括文本文档、PDF文件、电子邮件、视频、图像和音频文件，如图1-5所示。虽然处理起来更复杂，但非结构化数据通常提供更丰富的信息和见解，在机器学习和自然语言处理等领域尤其有价值。
- **半结构化数据**: 半结构化数据介于结构化数据和非结构化数据之间，它们可能不符合严格的数据库模型，但包含标签或其他标记来分隔语义元素，并使元素的层次结构可识别。XML和JSON是半结构化数据的典型例子，它们被广泛用于网络数据交互。
- **实时数据**: 实时数据是指需要立即处理的数据，以便快速做出响应或决策。这类数据在金融交易、网络分析、物联网（Internet of Things, IoT）设备监控和在线广告投放中非常常见。实时数据处理通常要求具有较高的技术能力，以便快速捕捉、分析和响应数据流。
- **时间序列数据**: 时间序列数据是按照时间顺序收集的数据信息，通常用于分析数据的趋势、周期性和季节性变化等，如图1-6所示。例如股票价格、气象记录和监控数据等都是时间序列数据的典型例子。
- **地理空间数据**: 地理空间数据含有关于地理位置的信息，这类数据在规划、物流和位置分析中非常关键。例如地图数据、卫星图像和GPS追踪数据都属于这一类型。
- **元数据**: 元数据是描述其他数据的数据，如图1-7所示，它可以包括文件大小、存储路径、创建日期、作者信息等。元数据有助于组织、管理和检索数据，通常是数据管理、数据仓库、数据湖中不可或缺的核心组成部分。



图 1-5

元数据

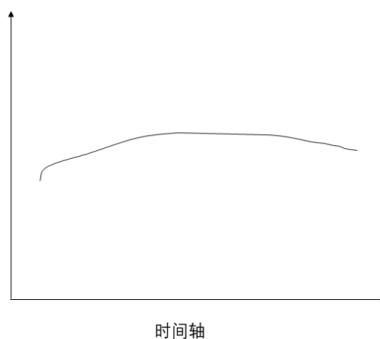


图 1-6

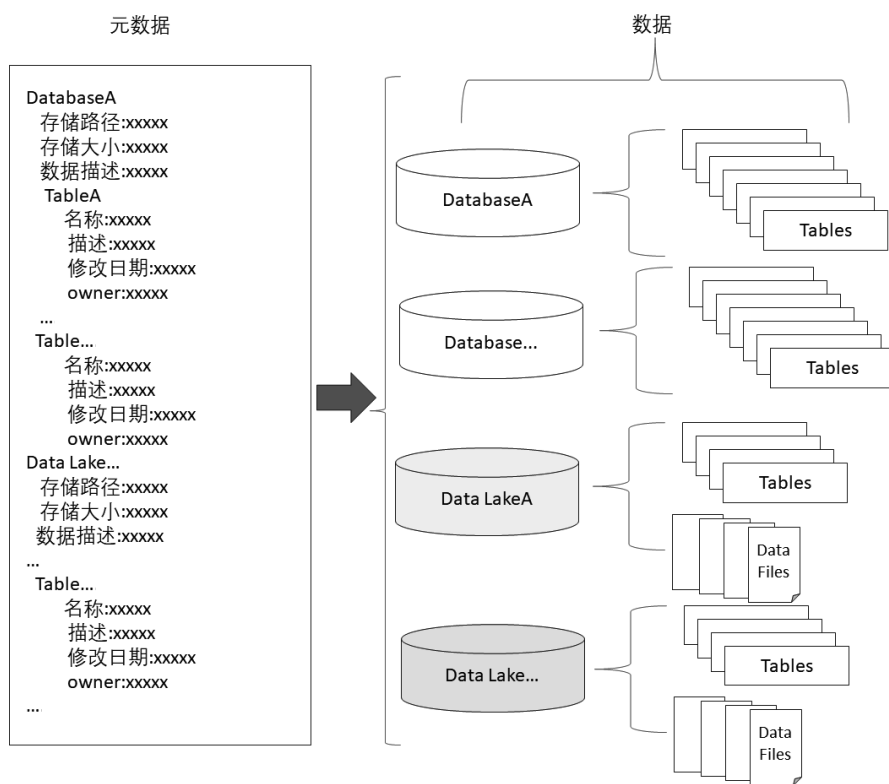


图 1-7

## 1.2 数据资产的分类

本节来看一下数据资产的分类方式。一般情况下，数据资产包含如下几种分类方式。

- 根据数据敏感性分类：根据数据的敏感程度，通常可以将数据分为不同的级别，如公开数据、内部使用数据、敏感数据、隐私数据、绝密数据等。这种分类方式有助于企业或者组织对不同级别的数据采取不同的保护措施，以确保数据的安全性和隐私性。

- 根据数据来源分类:根据数据的来源,通常可以将数据分为很多不同的类别,如会员数据、商品数据、业务数据、交易数据、第三方数据等。这种分类方式有助于企业或者组织更好地了解数据的来源和用途,从而更好地利用数据。
- 根据数据用途分类:根据数据的用途,通常可以将数据分为各种不同的类别,如分析数据、决策数据、销售数据、风控数据等。这种分类方式有助于企业或者组织更好地了解数据的应用场景和使用目的,从而更好地发挥数据的使用价值。
- 根据数据格式分类:根据数据的格式和类型,通常可以将数据分为更多不同的类别,如半结构化数据、结构化数据、非结构化数据、文本数据、图像数据、音频数据等。这种分类方式有助于企业或者组织更好地了解数据的结构和特点,从而更好地处理和保存数据。

## 1.3 数据资产的价值评估

### 1. 通过成本来评估数据资产的价值

利用成本来评估数据资产的价值是一种在数据领域经常使用的方法,主要通过考量数据的获取、处理、存储以及后期维护和升级的成本,来确定数据能够产生多大的价值。

- 获取成本:指的是获取自己想要的数据需要花费的成本,比如数据的采集成本(比如通过爬虫等方式采集)、购买成本(比如从第三方数据管理机构直接购买数据)等。获取成本中还应该包括数据的传输成本,因为无论是自己采集还是购买,数据都需要传输才能进入自己的管理系统中。
- 处理成本:指的是对指定的数据进行处理需要花费的成本,比如数据的清洗、转换和整合的成本。
- 存储成本:指的是在获取到数据后,经过数据处理,存储到指定存储介质中需要花费的成本,比如硬件和软件成本,以及维护和升级这些硬件和软件系统的成本。
- 维护成本:指的是对数据进行维护花费的成本,比如对数据进行更新、修正以确保数据的准确性和完整性等发生的成本。
- 升级成本:指的是对数据进行升级需要花费的成本,比如进行技术手段上的升级以确保数据更加及时和准确等发生的成本。

### 2. 通过收益来评估数据资产的价值

利用收益来评估数据资产的价值是指基于现有的数据资产在过去的应用和使用情况以及未来的应用场景来评估数据资产能产生多大的价值。该方式的评估步骤如图1-8所示。

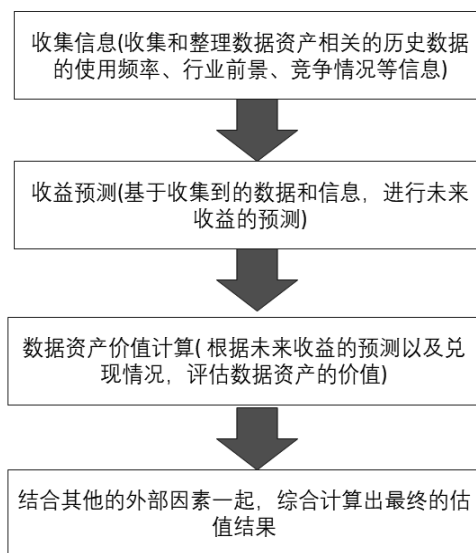


图 1-8

从图1-8中可以看到，利用收益来评估数据资产价值的核心点在于对行业相关的历史数据进行深入分析，并考量其未来前景以及竞争力。

## 1.4 数据资产的质量

确保数据高质量是数据资产管理的核心之一，企业或者组织管理其数据是因为需要使用数据或者挖掘数据中更大的价值，为了确保数据满足使用的需要，一定要做好数据质量的管理。如果数据质量过差，对于任何企业或者组织来说都是一种高成本的消耗。低质量的数据通常会产生如下不必要的成本开支，比如：

- 数据经常需要不断地返工和修正。
- 数据的质量低，导致企业或者组织的决策错误，从而造成巨大的经济损失。
- 数据的质量低，导致数据的使用变少，从而无法让数据发挥出应用的价值。

高质量的数据带来的相应好处包括：

- 可以更好地改善客户的使用体验。
- 可以更好地提升生产力。
- 降低低质量的数据造成的不可控风险。
- 高质量的数据可以带来更大的机会和机遇。
- 从对客户、产品、流程和机会的洞察中获得更大的竞争优势。

## 1.5 数据资产的存储

从传统的IT时代到现今的互联网时代和大数据时代，随着技术不断快速发展，数据资产的存储方式也发生着翻天覆地的变化。数据资产存储的发展历程主要分为以下几个阶段。

### 1. 文档存储时代

在信息化发展的早期，由于人们对数据的认识较少以及对数据价值的重要性认识不够，再加上当时IT系统的发展较为缓慢，数据主要依靠文档的方式存储到计算机上，如图1-9所示，比如通过Excel表格等方式来存储和查看数据。文档存储只能存储一些重要的数据，而且数据量不能太大。

通过文档存储数据的方式通常存在以下不足：

- 数据写入和修改的速度较慢，并且需要人手工把数据录入文档中，效率非常低下。
- 数据管理和维护较难，由于都是手工管理，因此极易出错，并且需要花费大量的时间，人力成本非常高。
- 数据不方便查看和检索，由于数据是以文档的方式存储的，当查找多个文档中的数据时，需要人工手动到每个文档中进行查找。

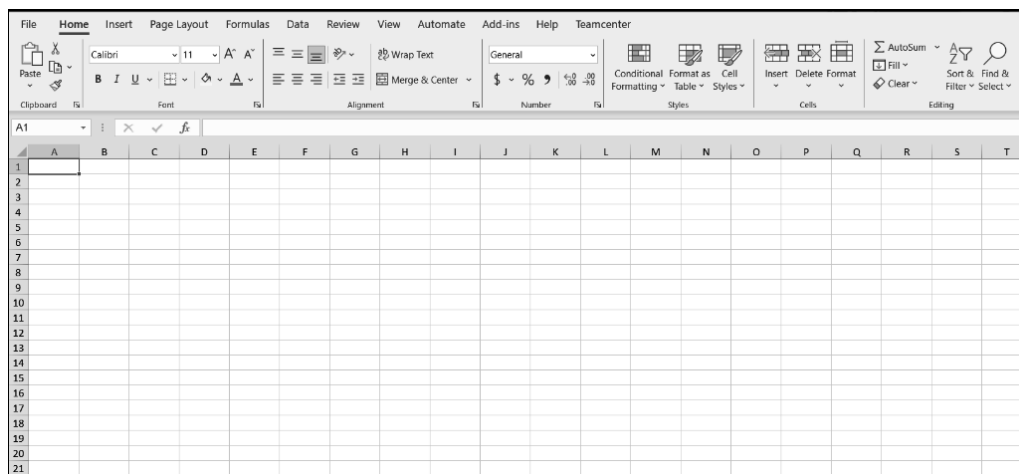


图 1-9

- 数据与数据之间的关联关系很难在文档数据库中记录，导致只能对数据进行一些简单的分析，无法进行复杂的关联分析。

## 2. 普通数据库存储的时代

如图1-10所示，随着计算机性能的更新换代以及IT技术的发展，开始出现了数据库技术，并且随着数据库技术的成熟发展以及SQL Server、MySQL、Oracle等很多关系数据库的出现，人们开始将数据存入数据库中。数据库的出现标志着数据资产信息化进程取得重要进展。



图 1-10

数据库存储可以解决很多文档存储数据的不足，比如不再需要人工录入和修改数据，可以通过数据库检索进一步提高数据的查询效率，通过数据库管理可以减少人工的手动管理和维护操作，数据库还可以存储数据与数据之间的关联关系，这样通过数据库就可以进行更复杂的数据分析以及数据查询等操作。

## 3. 数据仓库存储的时代

随着大数据的发展以及谷歌等大型科技公司对大数据技术发展的推动，人们对数据的认知进一步加深，对数据价值的探索不断加大，信息化技术的发展推动小数据逐步进入大数据时代。在谷歌三大核心论文的推动下，开源社区涌现出了很多以Hadoop、HBase、HDFS、Hive为首的优秀的大数据开源项目。正是在这个时期，人们开始提出数据仓库的概念。在大数据时代，随着人们对数据的需求越来越大，数据存储的体积也在急速膨胀，普通的数据库已经无法存储海量的数据了，更加无法对海量的数据进行分析了。数据仓库的引入通常可以解决如下问题。

- 解决数据分散的问题：如图1-11所示，在很多企业中都存在多套不同的业务系统，每个业务系统负责不同的业务，并且每个业务系统的数据通常都是存储在各自的数据库中，这样就会让数据非常分散。在数据仓库中，数据会采用集中式的存储架构，会将所有数据汇集到一个中心化的存储平台中，从而方便数据的整合和处理，以及进行更深入的数据挖掘和分析，让数据产生更大的价值。

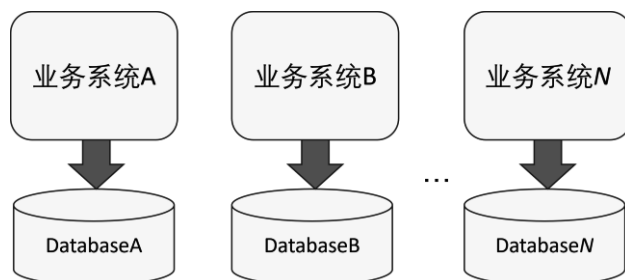


图 1-11

- 实现数据的标准化：通常来说，在一个企业或者组织中，不同的部门或者业务领域的数据标准和格式可能是不一样的，这就会给数据的整合和分析带来很多不便。有了数据仓库后，企业或者组织就可以将数据按照统一的标准进行转换和存储，从而实现数据的标准化。
- 保护数据的安全：随着人们对隐私和安全日益重视，保护数据的安全和隐私显得越来越重要。在数据仓库中，可以通过统一的数据加密、脱敏和权限访问控制来保护数据的安全和隐私。另外，随着大数据技术的发展，数据仓库的底层都是通过分布式的文件存储系统来存储数据的，从而让数据的完整性更加有保障。
- 更好地进行数字化转型：在数字化时代下，很多传统的企业或者组织都在不断进行数字化转型以适应市场的快速变化和提升自身产品的竞争力，数据仓库的出现为企业或者组织提供更加全面和完整的数据支持，帮助企业或者组织更好地了解市场的动向，快速响应客户的需求，从而制定适合自己的数字化转型策略。

#### 4. 数据湖存储的时代

随着大数据技术的快速更新换代以及数据湖概念的提出，以Databricks为首的大数据科技公司推出了名为Delta Lake的数据湖项目，在开源社区也出现了Hudi、Iceberg等优秀的数据湖项目。数据湖存储的引入，弥补了数据仓库所缺乏的某些功能，比如：

- 数据仓库中只能存储结构化的数据，而在数据湖中则没有任何限制，数据湖中不但可以存储结构化的数据，还可以存储半结构化甚至非结构化的数据。
- 数据仓库一般用于存储处理后的数据，而数据湖既可以存储没有经过处理的原始数据，也可以存储处理后的数据。
- 在技术层面，数据湖可以解决一些数据仓库无法解决的技术难题，比如以Hive为首的数据仓库在数据更新和数据删除等方面的能力非常弱，并且无法支持像数据库一样的事务处理。而数据湖则解决了这些难题，在数据湖中，可以高效地进行数据的更新和删除操作，并且支持事务处理，允许数据处理失败时执行回滚操作。

#### 5. 湖仓一体的时代

湖仓一体这个概念是最近几年才慢慢被提出来的，湖仓一体是将数据仓库和数据湖的优势结合起来而发展出来的一种全新的数据处理和存储架构。湖仓一体架构可以将结构化、非结构化的数据统一存放在一个共享的存储平台中，并且支持多种类型的数据源的接入，以及不同类型的数据分析，从而能更有效地挖掘出数据的价值。



## 1.6 数据资产管理

数据资产管理是一个涉及数据识别、分类、存储、保护和使用的复杂过程。数据资产的管理包括需要知晓数据的来源、存储位置、质量、适用的合规要求以及如何最大限度地利用这些数据等。在进行数据资产管理时，还需要注意数据的生命周期。数据的生命周期通常包括数据的创建、存储、使用、共享、归档和销毁等阶段。在数据的生命周期每个阶段，都需要相应的管理措施来保护数据的价值并确保其质量和合规性。

### 1. 数据资产管理的内容

通常来说，数据资产管理包括以下几个方面。

(1) 数据获取管理：通常指的是从数据源端获取数据的管理，比如当存在很多数据源时，需要对每个采集数据的数据源进行管理，如图1-12所示。

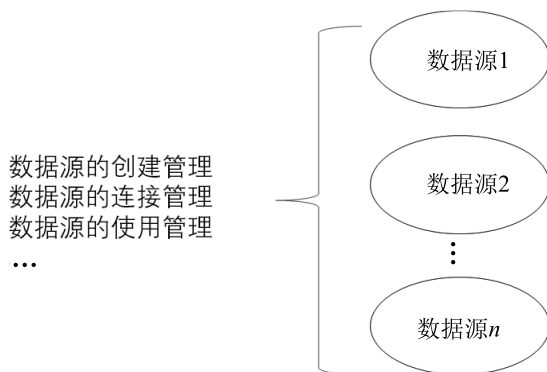


图 1-12

(2) 数据处理管理：当从数据的源端获取到数据后，通常还需要对数据进行一些加工和处理，比如数据格式的处理、数据的压缩处理、数据的异常值处理等，如图1-13所示。在大数据中，数据处理通常会使用专门的实时任务或者离线任务来处理，而数据处理管理通常需要对数据处理的任務进行管理，管理时需要知道数据处理中有没有报错、有没有处理失败等。

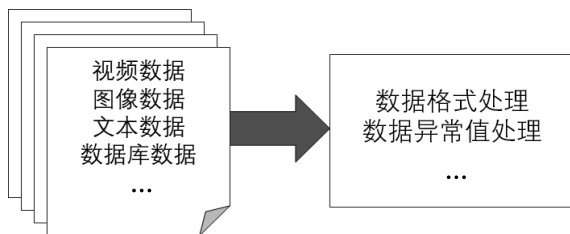


图 1-13

(3) 元数据管理：在前面已经提到，元数据是描述其他数据的数据，是数据资产管理的核心。如果没有元数据管理，用户在使用数据时就不知道数据是什么、数据包含什么信息、自

己需要的数据在哪里等，只有做好了元数据管理，才能让数据更容易被检索，才能让数据的使用者快速找到自己需要的数据。

(4) 主数据管理：是指对核心业务的实体相关的关键数据进行管理。在不同的企业或者不同的环境中，主数据可能是不同的。主数据管理可以进一步提高数据的价值，提升数据对业务的响应速度。

(5) 数据血缘管理：是指对数据之间的关联关系进行管理。通过数据血缘管理，数据的使用者就可以知道数据是从哪里来的、数据做了什么处理和加工等，如图1-14所示。

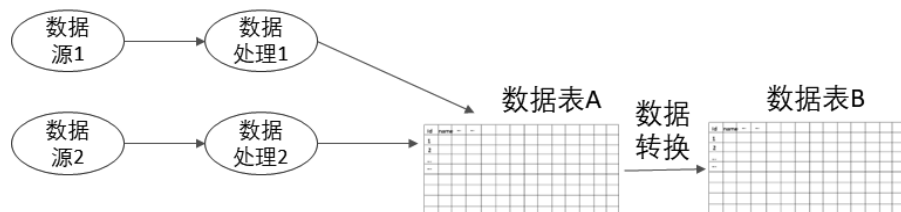


图 1-14

(6) 数据质量管理：通过对数据质量规则的定义来衡量数据的质量管理。数据质量的好坏直接关系数据的价值。

(7) 数据监控管理：数据监控管理是对数据链路、数据任务、数据服务、数据处理资源等环节进行监控与告警。当发现问题时，能够及时将问题告警和通知出来，以便数据的运维人员或者管理人员及时对数据进行处理。

(8) 数据服务管理：在数据资产中，数据服务是对外提供使用和访问的一种最重要的形式。数据只有对外提供访问，才能体现其自身的价值。数据服务的管理就是对这些对外提供数据访问所使用的服务进行管理。

(9) 数据权限与安全管理：在数据资产管理中，数据权限与安全的管理是让数据的整个生命周期中不会出现数据在未经授权的情况下被滥用，从而保护数据的安全和隐私不受侵犯。

## 2. 数据资产的管理方式

通常来说，数据资产的管理包括以下几种方式。

(1) 加强数据治理：通常来说，数据治理是做好数据资产管理的核心，通过不断建立和完善数据治理的流程和规范，明确数据管理的职责和分工，对数据做好分类和标记，让数据更方便地被查找。

(2) 建立完善的数据质量体系：数据质量直接决定了数据能否发挥其应有的作用，健全完善的数据质量体系可以持续不断地提高数据质量，让数据能够更准确地支撑企业或者组织的决策。

(3) 建立完善的数据权限和安全管理体系：数据安全是整个数据资产管理的基础。建立一套包括数据备份和恢复、数据加密和解密、数据权限控制等在内的体系，可以让数据更加安全可靠。

(4) 通过数据分析挖掘数据的更多价值：数据分析是数据资产的核心应用。在数据资产管理中，需要对数据进行更多分析，以挖掘出数据的更多潜在价值。

## 1.7 数据资产管理的信息化建设

数据资产管理的信息化建设，通常是指通过类似大数据等信息化技术对企业或者组织的数据资产进行管理和维护。本节讲解一下数据资产管理的信息化建设的好处和核心要素。

### 1. 数据资产管理信息化的好处

数据资产管理信息化可以带来如下好处。

- 及早发现数据问题：通过数据资产管理的信息化可以强化数据的质量以及监控和告警，当数据出现问题时，能够及早被发现。
- 提高数据管理的效率：通过大数据等IT技术手段，实现自动化、智能化管理数据，减少人工操作以及人为失误，降低人力成本和数据出错的风险。
- 让数据可以更快地产生价值：通过大数据等IT技术手段，让数据分析、数据挖掘更加迅速，能为企业或者组织提供更快、更准确的决策。
- 让数据可追溯和跟踪：通过建设数据资产管理平台，管理数据的处理过程和血缘关系等，让数据的使用者能对数据进行溯源。

### 2. 数据资产管理信息化建设的核心要素

数据资产管理信息化建设的核心要素如下：

- 数据采集：通过信息化的方式实现自动、实时、准确地在各个业务系统或者软硬件设备上采集数据，如图1-15所示。
- 数据处理：通过Spark、Flink等大数据技术，实时地对采集到的数据进行清洗和转换处理，挖掘出更多的数据价值。
- 数据存储：通过数据仓库或者数据湖等分布式存储的技术手段来存储不同数据种类和格式的海量数据。
- 数据服务：搭建统一的数据服务平台，让数据能够被业务需求轻松地访问到。
- 数据安全：建立信息化的安全机制，自动识别数据中可能存在的安全访问风险，对数据进行自动备份，以便在数据丢失时能够自动恢复。

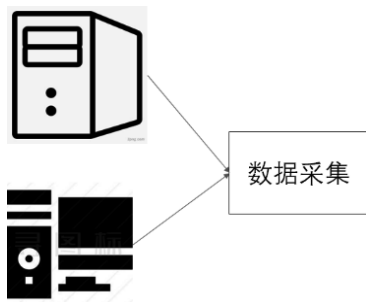


图 1-15

## 1.8 数据资产与人工智能

随着新一轮科技技术的变革，人工智能已经成为当前IT技术的热点话题之一，而数据更是赋能人工智能发展的关键。通常来说，数据与人工智能之间的关联关系可以通过图1-16来描述。

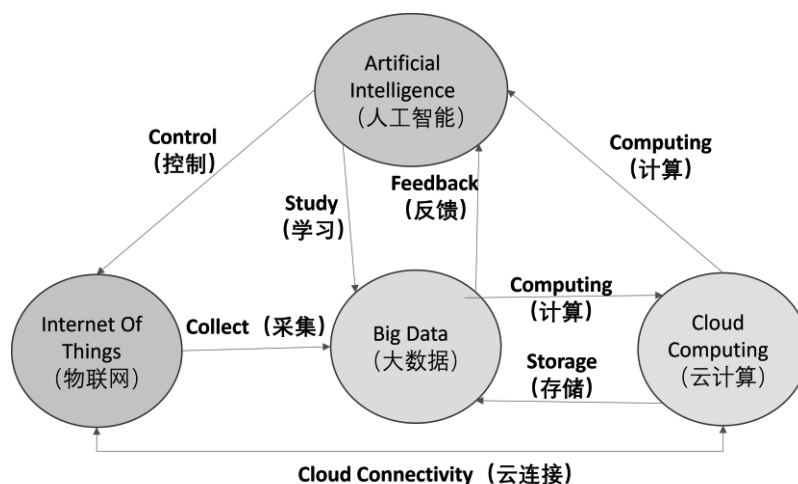


图 1-16

可以看到，人工智能在算法学习和模型训练时需要大量的数据做支撑，而人工智能算法预测的结果又需要通过数据反馈来验证其准确的程度，所以数据是支撑人工智能发展的关键，提高人工智能的准确性需要大量的数据来不断训练其模型。

数据资产和人工智能的结合将会使得：

- 人工智能更加智能化，能解放更多的人力成本。
- 能更好和更快地推动很多传统企业或者组织进行数字化转型。
- 加速科技发展的进程，发现更多未知的规律和现象。

总之，数据资产与人工智能的结合可以为很多企业或者组织带来更大的商机，可以让企业或者组织更好地理解数据，探索和挖掘数据中更多潜在的价值。