第一章　数据管理理论和方法

1.1. 数据管理的背景和技术发展史

1.1.1. 数据管理的背景

一、大数据全面纳入国家战略

近年来，移动互联网、云计算、人工智能、区块链技术等新一代数据技术迅猛发展，推动了人类社会从信息时代向数据时代的转变。在数据时代，新模式、新业态不断涌现，人类社会的生产生活方式正在发生深刻的变革，社会经济将出现一种全新的形态即数据经济，其将成为推动全球经济增长的重要驱动力。数据作为数据经济的关键要素，其作用是不容忽视的。正如麦肯锡在《Big Data: The Next Frontier For Innovation, Competition and Productivity》报告中所说，数据已成为不可或缺的，除劳动力和资本以外的又一重要的生产力要素。

发达国家非常重视数据经济的发展。美国已早于其他国家率先进入了数据时代，美国政府为了保持其领先地位制定了多个国家层面的战略措施，发布了《联邦云计算战略》，明确了“云计算”在国家政策中的战略地位，采取了“云优先”战略，发布了《大数据的研究和发展计划》，将大数据研究提升为事关国家核心竞争力的国家战略；英国颁布了《数字经济法》，推出了《数字大不列颠》的实施计划，出台了《英国数据能力发展战略规划》、《数字英国战略》与《数字经济战略》，将发展数据经济作为目标；

德国发布了《数字战略 2025》，为建设数据强国部署战略方向，规划国家战略层面上数字化转型的基本路径。

相较于国外，我国数字经济发展起步较晚。2015 年，党的十八届五中全会提出“实施国家大数据战略”。2016 年，在“十三五”规划纲要中进一步提出了如何实施发展大数据。2016 年 10 月 9 日，对于如何实施网络强国战略，中共中央组织并进行了集体学习。2017 年，党的十九大报告提出了要加速推动实体经济与大数据、互联网、人工智能深度紧密融合。2017 年 12 月在十九大后第二次集体学习中，习近平总书记强调数字经济关键要素在于数据，国家治理现代化水平的实现离不开大数据。2018 年 1 月 G20 阿根廷峰会上，习近平总书记提出要通过创新来促使实体经济和数字经济更好融合。2019 年 5 月 16 日发布了《数字乡村发展战略纲要》。2019 年国务院《政府工作报告》强调要促进区域协调发展，将区域一体化发展上升为国家战略。2020 年 4 月 9 日，国务院颁布了《关于构建更加完善的要素市场化配置体系机制的意见》文件，文件中指出需要抓紧发展数据要素市场的工作。所有相应的行为、措施和政策都充分表现了中央对于发展数字技术、数字经济的态度是坚定不移的。

表 1-1. 我国数字经济发展过程中的重要国家政策

年份	关键词	关键节点
2015	国家大数据战略	大数据完成顶层设计，上升为国家战略
2016	信息流；数据开放共享；国家大数据中心	行业大数据政策文件陆续出台
2017	数字中国	地方政府积极出台大数据相关政策
2018	与实体经济深度融合	地方政府陆续成立大数据中心
2019	数字经济；数字丝绸之路	数据融入到数字经济、数字治理等相关政策体系中

二、电网公司数据资源特点

电网企业数据涉及生产、经营、管理等多方面，可大致分为三类：一是电网运行和设备检测或监测数据；二是电网企业营销数据；三是电网企业管理数据(在本章第三小节中给出了详细说明)。随着电网企业信息化建设的不断推进，新型智能设备不断应用到电网生产各个环节，电网企业数据量、数据类型、数据来源等都有了巨大变化，数据量以几何级爆炸式速度增长，同时数据类型越来越复杂多样，数据累积速度已从 TB 级别上升到 PB 级别等。

随着大数据时代的不断发展，数据正成为电网企业的重要“资产”。电网数据蕴藏大量经济信息，基于时空电力数据、宏观经济数据、企业经营数据等，利用大数据技术构建电力经济指数，可以客观反映经济运行情况，辅助预测经济发展趋势，为制定宏观经济调控政策提供参考。当前，我国电网数据管理研究处于起步阶段，随着智能电网、新型电力设备与电力信息系统的更新换代，随之而来的是海量数据待筛选、处理和挖掘。电网企业将面临着数据的存储、共享融合、价值提升等方面带来的数据管理挑战。

1.1.2. 数据管理技术发展史

随着大数据技术的不断更新和迭代，数据管理技术得到了飞速的发展，相关概念如雨后春笋一般应运而生，如从最初单一存储数据的数据库到数据集市、数据仓库、数据湖、数据平台、数据中台等，在这一过程中，数据储量爆发式增长、数据管理能力全面提升，为厘清不同阶段数据管理的

内涵，本节对相关名词术语及内涵进行系统的剖析，便于全面认识数据管理发展史。

一、数据管理技术发展史

按照囊括的数据范围和层次，数据管理概念出现的次序依次为：数据库、数据集市、数据仓库、数据湖、数据平台和数据中台。

图 1-1. 数据管理技术发展史

(一) 数据库

数据库是“按照数据结构来组织、存储和管理数据的仓库”，是一个长期存储在计算机内的、有组织的、有共享的、统一管理的数据集合。可视为电子化的文件柜——存储电子文件的处所，用户可以对文件中的数据进行新增、查询、更新、删除等操作。

(二) 数据集市

数据集市是企业级数据仓库的一个子集，主要面向部门级业务，并且只面向某个特定的主题，按照多维的方式进行存储，包括定义维度需要计

算的指标维度的层次等，生成面向决策分析需求的数据立方体。

数据集市可以分为两种：一种是独立数据集市，这类数据集市有自己的源数据库和 ETL (Extract-Transform-Load)架构；另一种是非独立数据集市，这种数据集市没有自己的源系统，它的数据来自数据仓库。当用户或者应用程序不需要、不必要、不允许用到整个数据仓库的数据时，非独立数据集市就可以简单为用户提供一个数据仓库的子集。数据集市的应用场景更偏向于应对业务数据快速高效应用的需求，一般用于商业智能系统中探索式和交互式数据分析应用。

(三) 数据仓库

数据仓库最早在上世纪 80 年代由 IBM 提出，关于数据仓库的定义，目前最具影响力的是由美国 William H. Inmon 在其所著的书中给出的“数据仓库是支持管理决策过程的、面向主题的、集成的、随时间变化的持久的数据集合”。数据仓库本质也是一种数据库，相比于数据库，它有以下 4 个特点：面向主题，数据仓库中的数据是按照一定主题进行组织的，是决策时关心的重点。数据集成，数据仓库中的数据是从原始数据中合成而来，而原始数据互不相通。不可更新，数据仓库中的数据是为了决策分析使用，需要长期积累，常做的操作是查询，而非修改。随时间修改，数据仓库中的数据往往包含时间节点，是以时间为变量的动态发展过程，有助于发现业务流程中潜在的数据规律。

(四) 数据湖

数据湖(Data Lake)是 Pentaho 的 CTO James Dixon 提出来的，是一种数

据存储理念——即在系统或存储库中以自然格式存储数据的方法。以其自然格式存储数据的系统或存储库。数据湖通常是企业所有数据的单一存储，包括源系统数据的原始副本，以及用于报告、可视化、分析和机器学习等任务的转换数据。可以包括来自关系数据库(行和列)的结构化数据，半结构化数据(CSV、日志、XML、JSON)，非结构数据(电子邮件、文档、PDF)和二进制数据(图像、音频、视频)。

(五) 数据平台

随着云时代到来，企业所需资源数暴增。虽然“基础设施即服务”(Infrastructure as a Service，IaaS)类型供应商的出现从一定程度上解决了资源切割调度问题，但并未很好的解决基础设施资源与应用的融合。因此企业需要一种介于 IaaS 与“软件即服务”(Software as a Service，SaaS)之间的层级，用于屏蔽和控制 IaaS，快速开发和托管 SaaS，处于 IaaS 与 SaaS 之间的服务层为“平台即服务”(Platform as a Service，PaaS)。

数据平台处于 PaaS 层，属于“集成平台即服务”(integration Platform as a Service，iPaaS)。数据平台提供了数据接入、清洗、计算、存储、查询和分析的全流程自助化大数据服务。为数据运维人员提供一个低门槛的平台，根据平台提供的功能，快速构建面向大数据的可视化、智能化的运维支撑工具。数据平台不直接依赖于基础设施，并且数据的存取过程可表示为 SQL 语句查询。

(六) 数据中台

“中台”，是相对于前台和后台而生，是前台和后台的链接点，将业

务共同的工具和技术予以沉淀。数据中台是指数据采集交换、共享融合、组织处理、建模分析、管理治理和服务应用于一体的综合性数据能力平台，在大数据生态中处于承上启下的功能，提供面向数据应用支撑的底座能力。广义上来讲，数据中台是指通过企业内外部多源异构的数据采集、治理、建模、分析等应用，使数据对内优化管理提高业务，对外可以释放数据合作价值，成为企业数据资产管理中枢。数据中台建立后，会形成数据应用程序接口(Application Programming Interface，API)，为企业和客户提供高效的各种数据服务。

数据中台整体技术架构上采用云计算架构模式，将数据资源、计算资源、存储资源充分云化，并通过多租户技术进行资源打包整合，并进行开放，为用户提供“一站式”数据服务。数据中台利用大数据技术，对海量数据进行统一采集、计算、存储，并使用统一的数据规范进行管理，将企业内部所有数据统一处理形成标准化数据，挖掘出对企业最有价值的数据，构建企业数据资产库，提供一致的、高可用大数据服务。数据中台不是一套软件，也不是一个信息系统，而是一系列数据组件的集合，企业基于自身的信息化建设基础、数据基础以及业务特点对数据中台的能力进行定义，基于能力定义利用数据组件搭建自己的数据中台。

二、数据管理的提升

数据仓库作为数据行业发展时间轴上一以贯之的概念，它的存在见证了数据行业的发展，下面将以数据仓库为核心与其他五个概念的特性进行对比分析：

(一) 数据仓库与数据库

一般来说，传统数据库是为存储而生，而数据仓库很明显，是为分析而生。传统数据库包括增删改查，但数据仓库注重查询。传统数据库的主要任务是执行联机事务处理(OLTP)，主要负责日常操作。数据仓库系统在数据分析和决策方面为用户或“知识工人”提供服务，可以以不同的格式组织和提供数据，以便应付不同的需求，这种系统称作联机分析处理(OLAP)。

表 1-2. 数据仓库与数据库的对比

	数据仓库	数据库
面向对象	面向市场的、用于知识工人的数据分析	面向顾客的、用户操作员，客户和信息技术人员的事物和查询处理
数据内容	管理大量历史数据。提供汇总和聚集机制，而且在不同的粒度层上存储和管理信息	管理当前数据，一般这种数据比较琐碎，很难用于决策
数据设计	系统采用星形或雪花模型和面向主题的数据库设计	采用实体联系数据模型和面向应用的数据库设计
数据视图	经常需要跨越数据库模式的不同版本	关注一个企业或部分内部的当前数据，不涉及历史数据或不同单位的数据
访问模式	大部分是只读操作	一般需要并发控制和恢复机制

(二) 数据仓库与数据集市

数据集市不同于数据仓库，一般是服务于某几个部门。数据仓库向各个数据集市提供数据，且一般来讲，数据仓库的表设计符合规范化设计，而数据集市一般使用维度建模。一般有两种类型的数据集市——独立性和从属性。独立性数据集市直接从操作型环境获取数据，从属性数据集市从企业级数据仓库获取数据。数据仓库和数据集市的区别总结如下：

表 1-3. 数据仓库与数据集市的对比

	数据仓库	数据集市
数据来源	遗留系统、OLTP 系统、外部数据	数据仓库
范围	企业级	部门级或工作组级
数据粒度	最细的粒度	较粗的粒度
历史数据	大量的历史数据	适度的历史数据
优化	处理海量数据、数据探索	便于访问和分析、快速查询

(三) 数据仓库与数据湖

相较而言，数据湖是较新的技术，拥有不断演变的架构。数据湖存储任何形式(包括结构化和非结构化)和任何格式(包括文本、音频、视频和图像)的原始数据。数据湖在数据读取期间创建模式。与数据仓库相比，数据湖缺乏结构性，而且更灵活，它们还提供了更高的敏捷性。值得一提的是，数据湖非常适合使用机器学习和深度学习来执行各种任务，比如数据挖掘和数据分析，以及提取非结构化数据等。

表 1-4. 数据仓库与数据湖的对比

	数据仓库	数据湖
类型	结构化数据，而且这些数据必须与数据仓库事先定义的模型吻合	所有类型数据，如结构化数据、半结构化、非结构化数据等，数据的类型依赖于数据源系统的原始数据格式
目的	处理结构化数据，将他们转化为多维数据或报表，以满足后续的高级报表级数据分析需求	用于企业所有数据的单一存储，包括源系统数据的原始副本，以及用于报告、可视化、分析和机器学习等任务的转换数据
特点	高性能、可重复性、持续使用	便于探索、创新、灵活性高

(四) 数据仓库与数据平台

因数据仓库具有历史性，其中存储的数据大多是结构化数据，数据平台的出现解决了数据仓库不能处理非结构化数据和报表开发周期长的问题。

表 1-5. 数据仓库与数据平台的对比

	数据仓库	数据平台
数据类型	结构化数据	所有类型数据，如结构化数据、半结构化、非结构化数据等
服务方式	为业主提供服务的方式主要是分析报表	为业主提供的方式主要是直接提供数据集

(五) 数据仓库与数据中台

通过数据中台的定义我们可以总结出，数据中台是一站式解决平台，从数据集成、大数据计算、数据治理、数据工具、数据模型、数据应用、市场集成完整一套综合解决方案及产品系列。而数据仓库逐步从报表为主到分析为主、到预测为主、再到操作智能为目标。数据仓库系统的作用能实现跨业务条线、跨系统的数据整合，为管理分析和业务决策提供统一的数据支持。但数据中台从某个意义来说也属于数据仓库的一种，都是要把数据抽进来建立一个数据仓库。但是两者的数据来源和建立数据仓库的目标以及数据应用的方向都存在很大差异。

数据仓库也好，传统的数据平台也好，其出发点应该说更是一个支撑性的技术系统，即一定要去考虑系统有什么数据，然后系统才能干什么，因此特别强调数据质量和元数据管理，而数据中台的第一出发点可不是数据，而是业务，一开始不用看系统里面有什么数据，而是去解决业务问题需要什么样的数据服务。概括地说，二者的关键区别有以下几方面：