“数据中台”有何不同？丨建设数据中台系列（六）

* 耿立超

发布于：2020 年 8 月 6 日 11:17

回顾企业数据平台的发展史，我们可以梳理出三个重要阶段：

* 第一阶段：最初是以关系型数据库为基础的“数据仓库”，包括由此衍生出的商业智能（BI）和各类报表工具，这可以认为是第一代数据平台，它技术可靠、理论成熟，在数据体量不大、实时性要求不高的场景下有很好的适用性，多年发展积累下来的生态优势（包括技术、工具、人才等）使得它直到今天依然在很多企业中广泛应用着。
* 第二阶段：后来，随着大数据技术的崛起，数据平台在技术上有了新的选择和发展空间，特别是在实时数据处理和人工智能方面有了质的突破，与之相对应的是理论上的更新换代，人们从数据仓库理论延伸出了“数据湖”的设想，数据湖以大数据技术为支撑，在数据仓库理论上做了一些适当的扩充，提倡保留数据的原始形态以便为未来的各种分析留下空间，同时它更加倾向于使用 Schema on Read 而不是关系型数据库使用的 Schema on Write 进行数据处理。从理论上看，数据湖不足以像数据仓库一样成为一个独立的理论体系，人们更多的时候将它与大数据放在一起讨论，用来指代使用大数据技术构建的数据平台。
* 第三阶段：再之后就是随着中台架构一起提出的“数据中台”了。今天企业的 IT 生态可以自然地分为应用和数据两大部分，前者指的是各类业务应用系统，后者指的是数据仓库、数据湖等中心化的数据平台。由于这两类系统的设计思想和实现技术都有很大的差异，所以中台架构顺势分成了业务中台和数据中台两部分，这就是数据中台产生的大背景。

目前数据中台的主流架构都是以大数据技术为基础搭建的，具体来说，它是以 Hadoop 生态圈内的各种工具和 Spark 作为主要的技术堆栈，有的企业还会基于自身情况再搭配使用 Elasticsearch、MongoDB、Cassandra、Kudo 等其他数据处理和存储引擎。从技术架构上看，多以 Lambda 架构为主，划分批处理和实时处理两条数据处理通道。同时数据中台还往往具备人工智能及机器学习相关的能力，它会涵盖数据仓库和数据湖的全部职能，会提供强大的数据服务供业务中台和前台应用使用。本文核心观点援引自作者所著的[《大数据平台架构与原型实现：数据中台建设实战》](https://item.jd.com/12677623.html" \t "_blank)一书，全书对数据中台的理念、架构和具体实现做了详细论述。

1 数据中台所要解决的问题

在展开对数据中台的详实介绍之前，我们有必要对当前的企业数据生态做一些梳理，以便于我们能正对性地理解中台存在的意义。

目前，大多数企业特别是传统行业里的甲方企业，由于信息化水平的制约其数据生态还停留在较为落后的状态，存在着不少问题，以下是一些典型状况：

* 数据离散分布，信息孤岛的问题还没有完全得到解决；
* 依然在大量使用文件进行数据交互，没有实时 API，制约了上层业务流程的时效性；
* 企业数据处理平台依然依赖传统技术，负荷已经达到上限，无法进行水平伸缩；
* 大量数据离散于业务用户手工维护的文件中，难以自动化收集并处理；
* 同类型数据在多个业务系统中同步，数据冗余严重，一致性差，需要重复采集、核查、去重，成本高；
* 没有实时数据处理能力，无法快速及时地处理数据并反馈给业务用户；
* 没有健全的数据安全保障机制，面临数据泄露的风险；
* 缺乏完善的数据治理机制，数据质量参差不齐。

在过去，企业通过建立自己的数仓系统来统一存储和处理企业数据，这一过程要经历数年，并且会伴随着新数据源的产生持续进行。但是随着大数据时代的来临，传统数仓系统已经越来越难以支撑企业对数据处理的需求了，这体现在如下几个方面：

* 随着信息化的不断深入，企业产生的数据每年都在爆炸式的增长，传统数仓系统缺乏简单有效的水平伸缩能力，导致系统容量已经过饱和，系统性能遇到了“天花板”；
* 传统数仓系统只能处理关系型数据，很难处理非关系型数据；
* 企业对数据分析提出了更高的要求，在人工智能及机器学习等诸多新型领域都有迫切的需求，传统的数仓系统很难支撑这些新型的分析需求；
* 实时的数据分析越来越受到企业的重视，尤其是在一些大促等关键业务周期，传统数仓系统都是基于批量的离线处理，无法满足实时数据分析的需要。

这些因素都促使企业加大了对新一代数据平台的投入，通过构建基于大数据技术的新平台来应对新的挑战，而数据中台则成了目前大数据平台建设中最新、最热的方法论。

2 数据中台具备的能力

数据中台必须具备如下能力。

平滑自如的水平伸缩能力，从容应对海量数据

平滑自如的水平伸缩能力是数据中台必须具备的，特别是在数据体量迅速膨胀的今天，不具备存储和计算水平伸缩能力的平台是很难生存的，好在今天几乎所有的大数据技术都是分布式的，这赋予了数据中台天然的水平伸缩能力。

对资源拥有细粒度的控制能力，支持多任务、多用户下的作业处理

作为中心化的平台，企业不同部门和团队的数据都会存放在上面，每天会有大量的定时和即席作业运行，因此数据中台必须具备“多租户”的数据管理能力，对资源能进行细粒度的切分和调控。以 Hadoop 上的资源管理平台 Yarn 为例，通过定义各种动态资源分配策略，可以很好地协调各种作业之间的资源使用情况，确保各个业务线和不同用户的数据处理任务能及时有序地执行。

强大的实时处理能力

实时数据处理能力是以往传统数据平台所不具备的，这是数据中台的一大优势和亮点，通过实时处理我们可以将业务情况实时地反馈给用户，极大地缩短了业务用户的等待时间，提升了用户体验，在一些大促活动期间（如双 11），实时计算的时效性对于业务决策的支持作用会更加重要。

参与业务请求处理的能力

依托于实时计算能力，数据中台将有机会参与在线的业务处理，特别是在那些需要基于大量数据处理才能给出响应的业务请求（如用户积分的实时计算），过去这些处理都是通过批处理作业在夜间完成的，时效性和用户体验很差，现在通过数据中台可以实时地计算出结果并反馈给业务系统，这使得数据平台也开始参与在线的业务处理了。

具备人工智能及机器学习的数据分析能力

这是目前数据分析和应用领域最看重的能力，是当前数据分析领域的“皇冠”，它所带来的数据洞察能力是以往传统数据分析方法无法企及的，没有这种能力的数据中台是不完善的。这部分能力一般是通过在大数据平台上集成相关组件实现的（如 Spark MLib），但也有很多算法不能满足实际需要，因此需要集成一些第三方的算法库和集群环境作为补充。

以数据仓库理论管理和组织各类数据

数据仓库无疑是企业对于数据组织和管理的事实标准，不管是传统平台还是大数据平台，数据仓库理论都是科学有效的数据管理方法，可以说“没有数据仓库的大数据平台是没有灵魂的”。通过数据仓库体系的治理，企业数据的质量会得到大幅提升，也更利于前台的使用。

对外提供强大的数据服务，支持多种协议的数据传输与交互

过去的数据平台基本上都是将处理好的数据存放在关系型数据库中，供外围系统通过连接数据库的方式自行获取，可以说这是最低水平的数据服务，一个好的数据平台一定要提供强大的数据服务以便让数据需求方更容易和便捷地获取数据。平台支持的协议和方式越丰富，越能容易地帮助各业务中心和前台应用，加速集成和对接，降低企业整体的研发成本。而灵活便捷的数据获取方式又会吸引企业的数据供给方将数据主动放到数据中台上，从而享受数据中台带来的“红利”。

拥有完善的数据治理体系，数据质量能够得到有效保障

数据治理是贯穿数据平台建设全过程的一项工作，它是技术和管理方式的一种综合手段。数据中台一般会引入一些专业的数据治理工具对数据质量进行把控，这些工具会根据预定义的业务和技术规则定期抽检目标数据进行验证，并给出数据质量报告。为了配合数据治理，企业在管理上也应该成立相应的组织或机构来负责，这是建设数据中台在管理方面要做的工作之一。

精准的细粒度安全控制

数据中台要提供技术和管理上的多重机制保障企业的数据安全。从技术上看，数据中台需要提供严格的认证与授权机制来管理每一个使用平台的用户（包括自然人账户和应用系统账户），提供健全的数据加密与脱敏机制对敏感数据进行特殊处理，同时对每类数据的所有人、使用者和读写权限都要有明确的记录和追踪，对账户创建和授权申请都要有完备的审批机制。