02 | 关键抉择： 到底什么样的企业应该建数据中台?

2020-03-30 郭忆

数据中台实战课 进入课程 >

讲述：郭忆
时长 23:18 大小 21.35M

你好，我是郭忆。

在上一节课中，我和你一起回顾了大数据的发展历史，从历史脉络中，我们看到了数据中台凸显的价值，并得出数据中台是大数据下一站的结论。

既然数据中台受到了前所未有的关注，价值如此之大，是不是所有的企业都适合建设数据中台呢？到底什么样的企业应该建数据中台？带着这样的疑问，我们正式进入今天的课程。

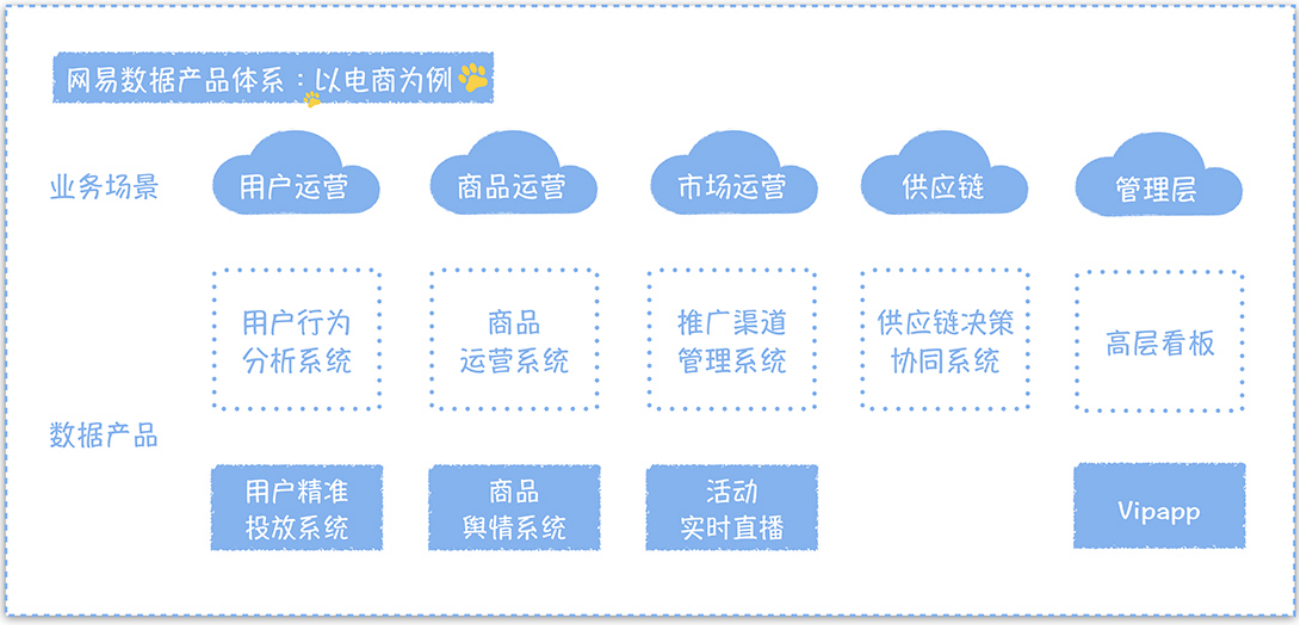
我先跟你分享一下，2018 年我们在建数据中台前面临的窘境，通过了解我们建数据中台的背景，你也可以对照着看一下自己所在的企业是否存在这样的问题，从而针对“是否需要构建一个数据中台”这个问题形成自己的看法。

建设中台前，我们面临的挑战

对于绝大多数互联网企业来说，2018 年绝对是煎熬的一年，因为面临线上流量枯竭，业绩增长乏力，企业成本高筑， 利润飞速下滑的风险。 原先粗放的企业管理模式和经营模式（比如我们在采购商品的时候，凭借经验去做出采购哪个商品的决策）已经没办法继续支撑企业的高速增长，越来越多的企业开始提数字化转型，强调数据是企业增长的新动力，它应该深入企业经营的各个环节。

数据需求的爆发式增长，促进了数据产品的蓬勃发展，在每个业务过程中，都有大量的数据产品辅助运营完成日常工作。例如，在电商的场景中，用户运营、商品运营、市场运营.....每个场景下，都有很多的数据产品，每天有大量的运营基于这些产品完成经营决策。

比如在供应链决策协同系统中，我们有一个智能补货的功能，会根据商品的库存、历史销售数据以及商品的舆情，智能计算商品的最佳采购计划，推送给运营审核，然后完成采购下单。



极客时间

大量数据产品的出现，在不断提高企业运营效率的同时，也暴露出很多尖锐的问题，在我看来，主要有五点。

1. 指标口径不一致。 两个数据产品一个包含税，一个不包含税，它们相同的一个指标名称都是销售额，结果却不一样。运营面对这些指标的时候，不知道指标的业务口径，很难去使用这些数据。

2. 数据重复建设，需求响应时间长。随着需求的增长，运营和分析师不断抱怨需求的交付时间拉长，面对快速变化的业务，需求响应时间已经无法满足业务对数据的敏捷研发要求。

3. 取数效率低。 面对数十万张表，我们的运营和分析师找数据、准确地理解数据非常困难，想找到一个想要的数据，确认这个数据和自己的需求匹配，他们往往需要花费三天以上的时间，对新人来说，这个时间会更长。

除了查找数据效率低，从系统中取数分析对于非技术出身的分析师和运营来说也是一个难题，这就导致大部分的取数工作还是依赖数据开发来完成。数据开发大部分的时间都被临时取数的需求占据，根本无法专注在数仓模型的构建和集市层数据的建设，最终形成了一个恶性循环，一方面是数据不完善，另一方面是数据开发忙于各种临时取数需求。

4. 数据质量差。数据经常因为 BUG 导致计算结果错误，最终导致错误的商业决策。**分享一个我们踩过的坑，**在大促期间，某类商品搜索转化率增长，于是我们给这个商品分配了更大的流量，可转化率增长的原因是数据计算错误，所以这部分流量也就浪费了，如果分配给其他的商品的话，可以多赚 200W 的营收。

5. 数据成本线性增长。数据成本随着需求的增长而线性增长，2017 年的时候，我们一个业务的大数据资源在 4000Core，但是 2018 就已经到达 9000Core 水平，如果折算成钱的话，已经多了 500 多万的机器成本。

相信你能在我“惨痛”的经历中，找到自己的影子，这些事儿的确很头疼，好在后来，我们用数据中台解决了这些问题。

为什么数据中台可以解决这些问题?

要想回答这个问题，你需要了解上述问题背后的原因。

指标口径不一致，可能原因包括三种：**业务口径不一致、计算逻辑不一致、数据来源不一致。**

如果是业务口径不一致，那就要明确区分两个指标不能使用相同的标识，像上面的例子，含税和不含税的两个指标， 不能同时叫销售额。

业务口径的描述往往是一段话，但是对于很多指标，涉及的计算逻辑非常复杂，仅仅一段话是描述不清楚的，此时，两个相同业务口径的指标，恰巧又分别是由两个数据开发去实现的，这样就有可能造成计算逻辑不一致。比如，有一个指标叫做排关单（排关单：把关单的排除；关单：关闭订单）的当天交易额这个指标，A 认为关单的定义是未发货前关闭的订单，B 认为关单是当天关闭的订单，大家对业务口径理解不一致，这样实现的计算结果也就会不一致。

最后，还可能是两个指标的数据来源不一样，比如一个来自实时数据，一个是来自离线的数据，即使加工逻辑一样，最终结果也可能不相同。

综合看来，要实现一致，就务必确保对同一个指标，只有一个业务口径，只加工一次，数据来源必须相同。

而数据需求响应慢在于烟囱式的开发模式，导致了大量重复逻辑代码的研发，比如同一份原始数据，两个任务都对原始数据进行清洗。如果只有一个任务清洗，产出一张明细表，另外一个任务直接引用这张表，就可以节省一个研发的清洗逻辑的开发。

所以，要解决数据需求响应慢，就必须解决数据复用的问题，要确保相同数据只加工一次，实现数据的共享。

取数效率低，一方面原因是找不到数据，另一方面原因可能是取不到数据。要解决找不到数据的问题，就必须要构建一个全局的企业数据资产目录，实现数据地图的功能，快速找到数据。而非技术人员并不适合用写 SQL 的方式来取数据，所以要解决取不到数据的问题，就要为他们提供可视化的查询平台，通过勾选一些参数，才更容易使用。

数据质量差的背后其实是数据问题很难被发现。我们经常是等到使用数据的人反馈投诉，才知道数据有问题。而数据的加工链路一般非常长，在我们的业务中，一个指标上游的所有链路加起来有 100 多个节点，这是很正常的事情。等到运营投诉再知道数据有问题就太迟了，因为要逐个去排查到底哪个任务有问题，然后再重跑这个任务以及所有下游链路上的每个任务，这样往往需要花费半天到一天的时间，最终导致故障恢复的效率很低，时间很长。

所以，要解决数据质量差，就要及时发现然后快速恢复数据问题。

最后一个是大数据的成本问题，它其实与需求响应慢背后的数据重复建设有关，因为重复开发任务的话，这些任务上线肯定会花费双倍的资源。如果我们可以节省一个任务的资源消耗，满足两个数据需求，就可以控制不必要的资源消耗。所以，成本问题背后也是数据重复建设的问题。

正当我们为这些问题苦恼的时候，数据中台的理念给了我们全新的启迪，那么数据中台到底是怎么一回事儿呢？**在我看来，数据中台是企业构建的标准的、安全的、统一的、共享的数据组织，通过数据服务化的方式支撑前端数据应用。**

数据中台消除了冗余数据，构建了企业级数据资产，提高了数据的共享能力，这与我们需要的能力不谋而合，所以很快，我们开启了数据中台的建设。

数据中台是如何解决这些问题的？

指标是数据加工的结果，要确保数据需求高质量的交付，首先是要管好指标。

原先指标的管理非常分散，没有全局统一的管理，在数据中台中，必须要有一个团队统一负责指标口径的管控。

其次，要实现指标体系化的管理，提高指标管理的效率。在指标系统中，我们会明确每个指标的业务口径，数据来源和计算逻辑，同时会按照类似数仓主题域的方式进行管理。

最后，要确保所有的数据产品、报表都引用指标系统的口径定义。当运营把鼠标 Hover 到某个指标上时，就可以浮现出该指标的口径定义。

通过对全局指标的梳理，我们实现了 100% 的数据产品的指标口径统一，消除了数据产品中，指标口径二义性的问题，同时还提供了方便分析师、运营查询的指标管理系统。

那么数据中台是怎么实现所有数据只加工一次的呢？简单来说，就是对于数仓数据，我们要求相同粒度的度量或者指标只加工一次，构建全局一致的公共维表。要实现上述目标，需要两个工具产品：

一个是数仓设计中心，在模型设计阶段，强制相同聚合粒度的模型，度量不能重复。

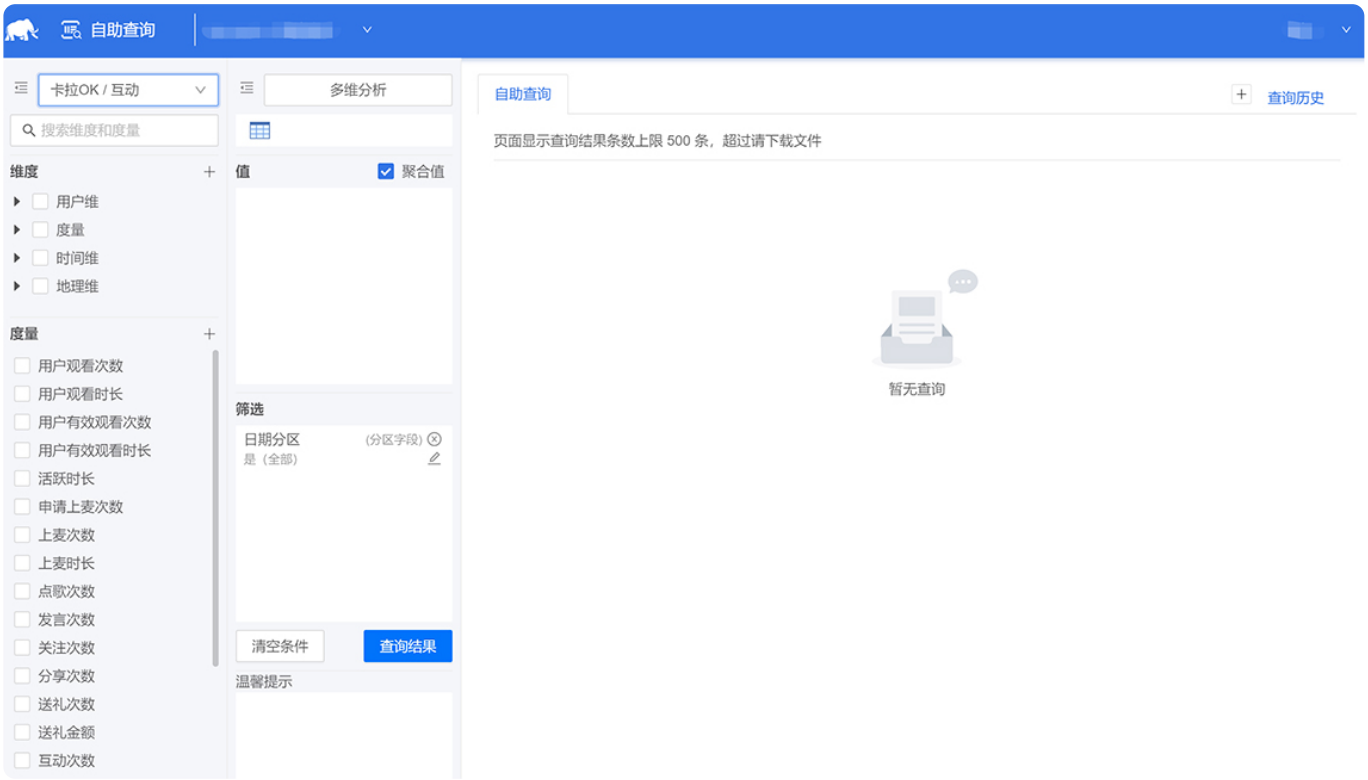
另外一个是数据地图，方便数据开发能够快速地理解一张表的准确含义。

这样就解决了数据重复加工导致研发效率瓶颈的问题，现在我们把需求的平均交付时间从一周减少到 2～3 天，大幅提高了数据产能，得到了分析师和运营的认可。

数据中台通过服务化的方式，提高了数据应用接入和管理的效率。原先数仓提供给应用的访问方式是直接提供表，应用开发自己把数据导出到一个查询引擎上，然后去访问查询引擎。在数据中台中，数仓的数据是通过 API 接口的方式提供给数据应用，数据应用不用关心底层不同的查询引擎访问方式的差异。

对于非技术人员，数据中台提供了可视化的取数平台，你只需要选取指标、通过获取指标系统中每个指标的可分析维度，然后勾选，添加筛选过滤条件，点击查询，就可以获取数据。

同时，数据中台构建了企业数据地图，你可以很方便地检索有哪些数据，它们在哪些表中，又关联了哪些指标和维度。通过自助取数平台和数据地图，公司的非技术人员开始自助完成取数，相比通过提需求给技术人员的方式，取数效率提高了 300%。



EasyFetch 网易自助取数界面

数据中台由于数据只能加工一次，强调数据的复用性，这就对数据的质量提出了更高的要求。而我们实现了全链路的数据质量稽核监控，对一个指标的产出上游链路中涉及的每个表，都实现了数据一致性、完整性、正确性和及时性的监控，确保在第一时间发现、恢复、通知数据问题。

原先，当技术人员问我们“今天数据有没有问题？” 的时候，我们很难回答，现在我们可以很自信地回答，数据对不对，哪些数据不对，什么时候能恢复了。我个人认为这个能力对我们最终达成 99.8% 数据 SLA 至关重要。

最后一个是成本问题。我们在构建数据中台的时候，研发了一个数据成本治理系统，从应用维度、表维度、任务的维度、文件的维度进行全面的治理。 从应用的维度，如果一个报表 30 天内没有访问，这个报表的产出价值就是低的，然后结合这个报表产出的所有上游表以及上游表的产出任务，我们可以计算加工这张表的成本，有了价值和成本，我们就能计算 ROI，根据 ROI 就可以实现将低价值的报表下线的功能。通过综合治理，最终我们在一个业务中节省了超过 20% 的成本，约 900W。

通过数据中台，最终我们成功解决了面临的问题，大幅提高了数据研发的效率、质量，降低了数据的成本。那么现在让我们回到课程开始时的问题，到底什么样的企业适合建数据中台？ 是不是所有企业都要构建一个数据中台?

什么样的企业适合建数据中台?

不可否认，数据中台的构建需要非常大的投入：一方面数据中台的建设离不开系统支撑，研发系统需要投入大量的人力，而这些系统是否能够匹配中台建设的需求，还需要持续打磨。另外一方面，面对大量的数据需求，要花费额外的人力去做数据模型的重构，也需要下定决心。

所以数据中台的建设，需要结合企业的现状，根据需要进行选择。我认为企业在选择数据中台的时候，应该考虑这样几个因素。

企业是否有大量的数据应用场景： 数据中台本身并不能直接产生业务价值，数据中台的本质是支撑快速地孵化数据应用。所以当你的企业有较多数据应用的场景时（一般有 3 个以上就可以考虑），就像我在课程开始时提到电商中有各种各样的数据应用场景，此时你要考虑构建一个数据中台。

经过了快速的信息化建设，企业存在较多的业务数据的孤岛，需要整合各个业务系统的数据，进行关联的分析，此时，你需要构建一个数据中台。比如在我们做电商的初期，仓储、供应链、市场运营都是独立的数据仓库，当时数据分析的时候，往往跨了很多数据系统，为了消除这些数据孤岛，就必须要构建一个数据中台。

当你的团队正在面临效率、质量和成本的苦恼时，面对大量的开发，却不知道如何提高效能，数据经常出问题而束手无策，老板还要求你控制数据的成本，这个时候，数据中台可以帮助你。

当你所在的企业面临经营困难，需要通过数据实现精益运营，提高企业的运营效率的时候，你需要构建一个数据中台，同时结合可视化的 BI 数据产品，实现数据从应用到中台的完整构建，在我的接触中，这种类型往往出现在传统企业中。

企业规模也是必须要考虑的一个因素，数据中台因为投入大，收益偏长线，所以更适合业务相对稳定的大公司，并不适合初创型的小公司。

如果你的公司有这样几个特征，不要怀疑，把数据中台提上日程吧。

课堂总结

本节课，我结合自己的经历，带你了解了企业数据在日常使用过程中面临的一些难题，通过分析，我们发现，数据中台恰好可以对症下药，解决这些问题。在这个过程中，我想强调这样几个重点：

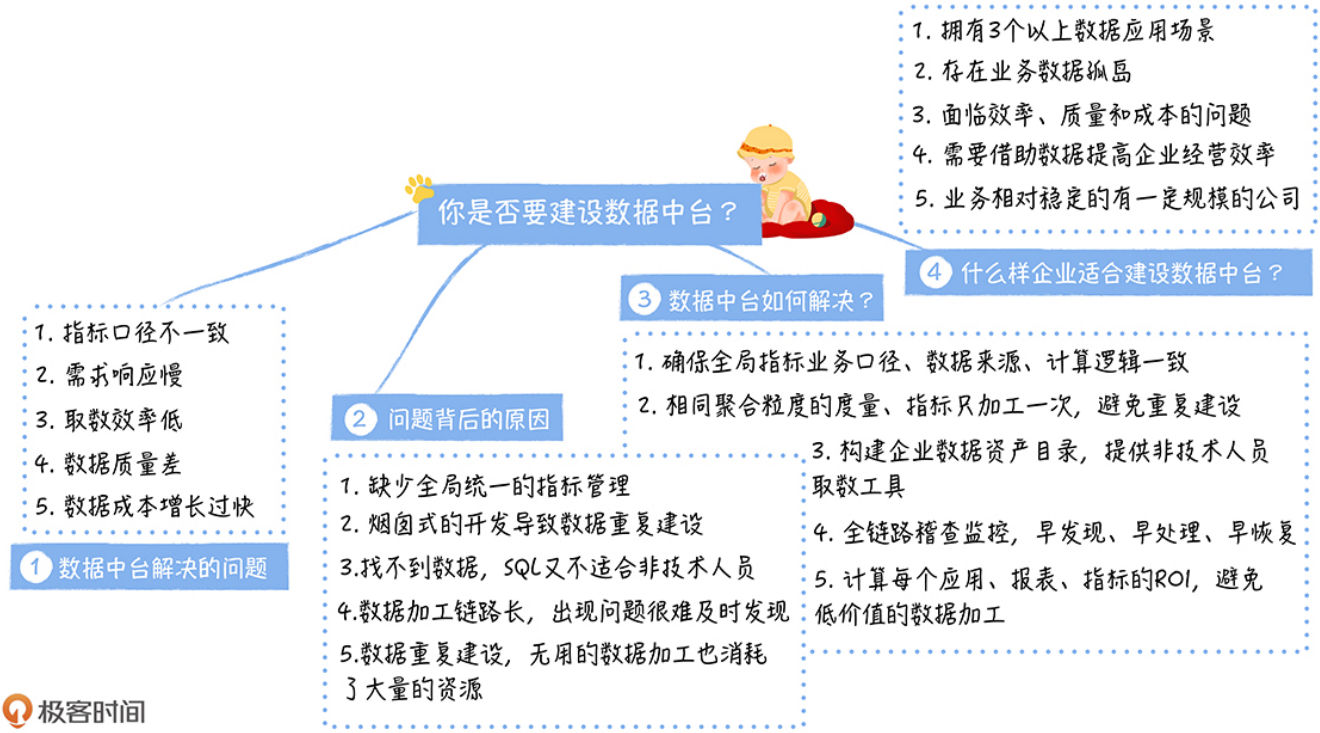
效率、质量和成本是决定数据能否支撑好业务的关键，构建数据中台的目标就是要实现高效率、高质量、低成本。

数据只加工一次是建设数据中台的核心，本质上是要实现公共计算逻辑的下沉和复用。

如果你的企业拥有 3 个以上的数据应用场景，数据产品还在不断研发和更新，你必须要认真考虑建设数据中台。

在最后，我想再次强调一下，建设数据中台不能盲目跟风，因为它不一定适合你，我在生活中见到了很多不符合上述特征，却想要建设数据中台的公司，比如一些初创型的小公司，初期投入了大量人力成本建设数据中台，因为业务变化快，缺少深入数据应用场景，结果却是虎头蛇尾，价值无法落地。所以，你最正确的做法是仔细想想我提出的上述 5 点要素。

因为这节课信息比较密集，我用一个脑图帮你梳理一下知识体系，便于你理解：



思考时间

我同样给留给你一道思考题，一个企业是不是只能建设一个数据中台？

最后，感谢你的阅读，如果这篇文章让你有所收获，也欢迎你将它分享给更多的朋友。

精选留言 (17)

写留言

Max 置顶
2020-03-30

如果业务场景分散。
或许一个中台就不太合适了。
还是应该适配发展第一。

展开

作者回复: 说到点子上了。

我就网易来举个例子，大家都知道，网易的业务涉及的领域非常的多，云音乐属于在线娱乐，有道是教育，严选、考拉（已经脱离网易）是电商，新闻属于传媒，我们有没有必要构建一个属于网易的数据中台?

在我看来，没有这个必要，因为业务之间相差很大，重叠很少，需要跨业务进行分析的场景很少。毕竟跨业务线的中台构建，需要更大的投入和成本，这样收益和投入不成比例，也没这个必要。

我非常认可这位同学的观点， 不管是组织，还是数据中台，都是为了更好的支撑业务，如果业务线比较独立，那确实会存在多个面向不同业务线的多个数据中台。

3

赵澈
2020-04-01

我是一家金融机构统计条线的员工，我们现在就打算开发数据中台，但是我有一个疑问，我们在构建公共维表的过程后，如何通过公共维表更好的搭建中台指标，同时利用什么方式可以对中台指标进行更好的展示?

1

幸运草
2020-03-30

一个数据中台应该是一个供应链全流程的数据吧，这些数据进行聚合产生价值，如果不属于同一个供应链，应用价值，数据模型，数据关联性几乎没有，没必要弄在一个数据中

台，所以一个企业的业务是不同的供应链，应该建不同的数据中台。个人观点，多多指教。

展开

作者回复: 这个回答不太准确。

首先我不是很确定你说的供应链是不是电商业务中的供应链。数据中台应该是面向一个业务线的，比如电商业务，教育业务，这些都是不同的业务线，应该有不同的数据中台，业务之间没有重叠，差别很大。

1

Miles
2020-04-01

感觉跟传统模式的数仓没太多区别，主要在数据应用层面会更加丰富吧

羽轩
2020-04-01

个人认为，作为互联网类创业公司，尤其是2B的SAAS类，初期也需要有中台思维来架构产品，甚至可以做一个V1.0版本的小中台，把底层架构打扎实，后续无论是转型还是扩大规模，都能做到心中不慌。
所以，关键是思维方式。

展开

亮
2020-04-01

一个好的中台建设需要大量人力，不同团队共同协作，有这个魄力的老板还挺难得

XiangJiawei
2020-03-31

老师，想请教如下问题，谢谢：
1）API的方式保证数据服务的执行效率？会不会导致所有的业务压力全部都压在数据中台上，导致API的返回速度有时候快、有时候慢，影响前前端的体验。
2）API的返回数据量是否有限制？如果有大批量的数据需要通过API获取是否合适，例如超过100MB甚至更大，这种场景如何处理？...

展开

邓子武
2020-03-31

数据中台是在做产品，传统方式是做一个个数据流和报表

作者回复: 这个观点不太准确，数据中台强调数据、接口的复用，实现数据的共享，从而提高效率、质量，节省成本，可以更好的支撑数据应用。

至于是产品还是报表， 数据中台之上，既有数据产品，也有BI报表。

张勇
2020-03-31

数据中台说白了就是数据平台数据大锅烩，从头到尾一个人做了,只提供API给上层.

作者回复: 这个理解不是特别准确哈。 首先数据中台是数据平台数据的大杂烩，这个就不太准确。数据中台强调数据的规范化建设，数据需要按照主题域、分层的模式，构建统一、共享、标准、安全的数据体系。数据平台你可以理解为是一个面向数据研发场景的工具集合。

至于从头到尾一个人做，就更不准确啦，数据中台的团队，涉及的人员非常多，在ODS层有一个小团队，然后DWD，DWS，DM 基本是按照主题域来管理的，ADS 又是独立的团队来管理。不存在一个人构建一个数据中台的情况哈。

Jxin
2020-03-31

1.数据平台解决基础数据的去重和可复用的问题。关键在同域多元数据的抽象整合。
2.数据中台解决基于数据平台搭建的，上层业务逻辑的去重复用的问题。关键在于多个应用域间，重叠的问题域的识别和抽象。

3.以上是以往个人理解，感觉跟栏主这个专栏应该能再刷新下认知了。...
展开

作者回复: 我的答案已经置顶，欢迎大家讨论~

小美
2020-03-31

衡量报表价值的维度都有那些呢？感觉这个很难量化，成本比较好计算。

作者回复: 数据应用分为两种，一种是报表展示类的，一种是面向特定应用场景的数据应用。你提到的是报表，所以我理解你是指第一种。

对于报表的话，我们衡量价值，一般是从报表的使用范围，有多少人看你这个报表，一般用一周内的周活。当然，这个人群要进行加权，比如管理级别的加权，老板的加权系数应该非常高。然后是报表的使用频率，一般用报表的PV来衡量。对于使用人数多，使用频率高的，我们可以认为，这类报表的价值越大。对于周活、PV 都是可以量化横向对比的。

我在第八讲精细化成本管理，会详细的再展开讲，请继续阅读哈。

西蒙
2020-03-31

希望老师可以分享下数据质量建设的整个过程，谢谢。

展开

作者回复: 别急哈，在第7讲，数据质量实现篇，我会详细的介绍数据质量问题的根源，治理方法。慢慢看哈~

leslie
2020-03-31

其实老师在提出数据中台是否能只建立一套的同时我想反问老师一个问题：一个企业是否只能有一套DevOps或者SRE？答案尽在其中

展开

作者回复: 我的答案已经置顶，欢迎讨论~

枫雪无痕
2020-03-31

数据中台强调共享，复用，标准统一，一个就够。多业务线，宏观统一管理，细分处理。

作者回复: 我的答案已经置顶，欢迎讨论哈~

coyang

2020-03-31

请问一下老师，你说的99.8%数据SLA，具体是指什么？是怎么算出来的？

作者回复: 数据SLA，我们会为数据中台的表设置一个基线产出时间，一般中台核心表我们要求是在6点半之前要产出，如果在6点半之前无法产出，我们就开始计算时间，每超过一分钟，就算一分钟不可用时间，99.8%，代表一年有18个小时，这个挑战还挺大的。

阿巍-豆夫

2020-03-30

数据地图，长的什么样子的？ 只闻概念，不了其实，能否有个简单的例子？ 还有指标和纬度是什么区别？ 什么才算指标？

展开

作者回复:
别急，我们在第4讲元数据管理部分，会讲数据地图，当时候你就知道数据地图的真面容啦。继续往下看。

指标，其实是事实，比如交易额、销售额、发货量等等。

维度，可以理解为环境，比如时间、地理、商品、卖家、买家等等。

1

Kǎfkā²⁰²⁰

2020-03-30

最好是一个，能最大化共享企业资源。不过，如果业务发展很快，来不及支持，暂时独立建设多个我觉得未尝不可，最终还是会走到一起的

展开

作者回复: 其实没有所谓的最好，只有最适合业务发展的组织模式和数据中台。

答案我已经置顶，欢迎讨论~