# 内容来源：《大数据分析师课程》

**大数据分析师职业守则：**

诚实公正、严谨求是、遵守章法、恪尽职守、勤勉好学、追求卓越。

**三次信息化浪潮：**

20世纪70-80年代：计算机普及，集成电路发展，晶体管取代了电子管，缩小了计算机体积；

90年代：互联网普及；

21世纪初：大数据技术发展，网络共享使大量数据积累下来，进入大数据时代，数据深度挖掘和融合应用

**大数据的发展：**

萌芽阶段 1980-2008，《第三次浪潮》这本书中第一次提到“大数据”，海量数据问题；

发展阶段：2009-2011，2009年联合国用大数据预测疾病，2011年麦肯锡发布关于“大数据”报告，首次定义大数据概念；

爆发阶段：2012-2016，2013年是中国大数据元年；

成熟阶段：2017-至今，数据安全和个人信息保护法，“十三五”规划发展。

**大数据带来的思维模式的改变：**

大数据思维：使用全部的样本数据，通过算法找出各种数据之间复杂的关系，而且这些数据并不要求完全准确，可以是混杂的，完全符合客观世界的真是规律。

数据思维：利用数据解决业务问题时，所表现出来的思维模式。

**大数据特点：5V**

Volume（大量）：KB、MB、GB、TB、PB

Variety（多样）：结构化，非结构化，半结构化

Value（价值密度低）：数据提纯

Velocity（高速）：增速快，处理速度快，处理时效性要求高

Veracity（真实性）：准确性，可信度，数据质量高

**大数据市场产业链**

大数据产业是以数据采集、交易、存储、加工、分析、服务为主的各类经济活动。包括数据资源建设，大数据软硬件产品的开发，销售，租赁活动和相关信息技术服务。

|  |  |  |
| --- | --- | --- |
| 产业链 | | |
| 上游基础支持 | 中游大数据服务 | 下游大数据应用 |
| 硬件设备：服务器、交换机、光模块 | 数据采集：互联网、物联网 | 智慧出行、医疗、政务、金融，工业 |
| 数据存储及云平台：存储硬件、数据中心（IDC）、云服务/IAAS |
| 基础设备：电力设备、基础软件、运营商、其他 | 数据分析或挖掘：软件&SAAS、基础分析工具 |
| 数据安全 |

# 内容来源：[大型集团企业数据治理实施的有效策略](https://mp.weixin.qq.com/s/IpsZ0BQlzyWBSz2ilu0KNw)

**一 概述**

随着世界越来越以数据为主导，大型组织面临着管理大量数据并确保其准确性、安全性和合规性的挑战。这就是数据治理的作用所在——创建和实施数据管理政策、程序和标准的过程。数据治理对于大型组织至关重要，因为它可以确保组织内的每个人都在数据管理方面达成共识，从而最大限度地降低数据泄露、合规性违规和其他数据相关问题的风险。

然而，实施有效的数据治理政策可能是一项艰巨的任务，尤其是在拥有复杂数据系统和多元化员工队伍的大型组织中。在本博客中，我们将探讨有效数据治理的关键组成部分，并提供在大型组织中实施数据政策的策略。我们还将研究成功实施数据治理的实际案例，并为希望改善数据治理的组织提供最佳实践和建议。

在当今的数字时代，数据已成为组织最宝贵的资产之一。大型组织每天都会产生大量数据，从客户信息到金融交易。这些数据对于做出明智的业务决策、推动增长和创新以及在市场上获得竞争优势至关重要。然而，随着数据量的增加，需要进行适当的管理，这就是数据治理发挥作用的地方。

数据治理是指组织为有效管理其数据资产而制定的政策、程序和标准。它涉及确保数据在整个生命周期（从创建到处置）的准确性、完整性、及时性和安全性。数据治理还建立了围绕数据管理的决策框架，确保组织以负责任和合乎道德的方式使用其数据。

对于大型组织而言，数据治理对于确保其数据的准确性、可靠性和安全性至关重要。它可以帮助组织管理跨部门、团队和利益相关者的数据管理复杂性。通过数据治理，组织可以建立一致的数据管理方法，使他们能够做出明智的决策并优化运营。

在接下来的部分中，我们将探讨有效数据治理的关键组成部分以及在大型组织中实施数据政策的策略。

二 有效数据治理的关键组成部分

有效的数据治理涉及一系列政策、程序和标准，它们共同确保在数据的整个生命周期中得到正确管理。以下是有效数据治理的一些关键组成部分：

数据管理员的角色和职责：数据管理员负责管理和监督组织内的数据。他们确保遵守数据政策、妥善管理数据，并且需要数据的人可以访问数据。为数据管理员建立明确的角色和职责对于有效的数据治理至关重要。

数据分类和元数据管理：数据分类涉及根据数据的敏感度、机密性和关键性对其进行分类。元数据管理涉及记录和管理数据属性，例如数据沿袭、数据质量和数据所有权。这些流程有助于确保根据组织政策和监管要求对数据进行正确的分类、管理和使用。

数据隐私和安全措施：数据治理还涉及保护数据免遭未经授权的访问、盗窃或滥用。这需要实施强大的数据隐私和安全措施，例如访问控制、加密和定期安全审计。

数据质量和完整性管理：确保数据的准确性、完整性和一致性对于有效决策至关重要。数据治理政策和流程必须包括维护数据质量和完整性的措施，例如数据验证检查、错误处理和数据清理。

数据生命周期管理：数据治理政策还必须解决整个数据生命周期，从数据创建到数据删除。这包括数据保留策略、数据归档和数据处置。有效的数据生命周期管理可确保在需要时保留数据，并以安全合规的方式处置数据。

三 实施有效数据政策的策略

实施有效的数据政策是一个关键过程，需要精心规划的方法和战略实施。以下是组织可以采用的一些关键策略，以确保成功实施有效的数据政策：

评估当前数据政策并找出差距：全面评估当前数据政策有助于找出现有流程中的差距和低效之处。此步骤有助于更好地了解组织的数据格局、数据质量以及遵守行业法规的水平。

制定全面的数据治理计划：全面的数据治理计划概述了管理整个组织如何收集、存储、管理和共享数据的政策、程序和标准。该计划应在整个组织的利益相关者（包括IT、法律、合规和业务部门）的意见下制定。

建立明确的角色和职责：成功的数据治理计划需要明确定义数据管理员、数据所有者和其他关键利益相关者的角色和职责。这些角色应该明确定义，并且应该对员工进行充分的培训和支持，以履行他们的职责。

确保利益相关者和高管层的支持：要成功实施数据政策，必须获得整个组织利益相关者（包括高管层）的支持。沟通和教育在确保数据治理计划的支持和支持方面可以发挥关键作用。

实施强大的数据质量和完整性控制：有效的数据治理计划需要强大的数据质量和完整性控制，以确保数据的准确性、完整性和一致性。数据质量控制应旨在识别和纠正发现的数据错误。

定期监测和评估数据政策的有效性：持续监测和评估数据政策有助于确定需要改进的领域，并确保数据政策仍然有效且相关。

为员工提供持续的培训和教育：有效的数据治理计划需要对员工进行持续的培训和教育，以确保他们了解政策和程序以及他们在数据治理计划中的角色和职责。定期培训可以帮助确保员工了解数据政策的变化并遵守行业法规。

四 案例研究：成功实施数据治理的示例

案例研究为大型组织中数据治理政策的有效性提供了实用的视角。以下是一些成功实施数据治理的示例：

宝洁：这家消费品巨头实施了全球数据治理框架，其中包括建立数据治理委员会、数据质量委员会和数据管理计划。这有助于他们在整个组织内标准化数据管理实践，从而提高数据的准确性、完整性和一致性。

世界银行：世界银行实施了全面的数据治理计划，以确保其数据高质量、安全并符合其战略目标。这包括建立明确的数据管理角色和职责、实施强有力的数据质量控制以及建立数据质量监测和报告系统。

沃尔玛：沃尔玛实施了数据治理框架，其中包括数据治理委员会、数据管理计划和数据质量管理计划。这有助于他们提高数据的准确性、一致性和完整性，从而做出更好的决策并提高运营效率。

强生公司：强生公司是全球最大的医疗保健公司之一，已实施了强大的数据治理计划，以确保遵守监管要求并保护患者数据。他们为数据管理建立了明确的角色和职责，并定期审查和更新其数据政策，以确保其保持相关性和有效性。

美国运通：作为一家处理敏感客户数据的金融服务公司，美国运通实施了全面的数据治理计划，以确保其数据的安全性和完整性。这包括定期数据质量评估、元数据管理以及数据隐私和安全措施。

经验教训和最佳做法：

数据治理政策应与组织的战略目标和宗旨保持一致。成功的数据治理实施需要强有力的执行领导和利益相关者的支持。

IT和业务利益相关者之间的有效沟通和协作至关重要。应定期审查和更新数据治理政策，以确保其保持相关性和有效性。

对员工进行数据治理政策和程序的持续培训和教育至关重要。

1.强生公司成功实施数据治理的案例

强生公司(J&J)是一家全球医疗保健公司，业务遍及60多个国家，拥有超过130,000名员工。作为一家受高度监管的敏感行业公司，强生公司认识到数据治理对于确保其数据资产的准确性和安全性的重要性。

强生公司实施了全面的数据治理计划，以确保其所有业务部门的数据准确、完整且值得信赖。强生数据治理计划的一些关键组成部分包括：

明确职责：强生成立了数据治理委员会，由各业务部门的高层领导和数据管理员组成。该委员会负责制定数据政策、建立标准和指南，并确保整个组织的合规性。

数据分类和元数据管理：强生公司创建了一个数据分类框架，根据数据的敏感性、关键性和监管要求对其进行分类。他们还建立了元数据管理流程，以确保准确标记和描述数据。

数据隐私和安全措施：强生实施了强大的安全控制措施来保护其敏感数据，包括加密、访问控制和监控工具。他们还制定了数据隐私政策，以确保遵守全球数据隐私法规。

数据质量和完整性管理：强生实施了数据质量控制和流程，以确保所有业务部门的数据准确、完整和一致。他们还建立了数据沿袭流程来跟踪整个组织的数据流动。

数据生命周期管理：强生公司建立了管理整个数据生命周期的流程，包括数据获取、存储、保留和处置。他们还建立了归档和检索历史数据的流程。

总体而言，强生的数据治理计划取得了成功，提高了数据质量、效率和整个组织的决策能力。强生数据治理实施的一些重要经验教训包括：

建立明确的数据治理角色和职责，并确保所有利益相关者都参与该过程。

制定全面的数据治理计划，涵盖数据管理的各个方面，从数据质量到数据安全。

确保整个组织的行政领导和利益相关者的认同，以确保该计划得到适当的资金和资源。

定期监控和评估数据政策和流程的有效性，以确保持续合规和改进。

为员工提供持续的培训和教育，以确保他们了解自己的角色和职责，并具备正确管理数据的能力。

2.跨国金融服务公司美国运通成功实施了一项数据治理计划，帮助其有效地管理数据并做出更明智的决策。

该公司认识到，其收集的大量数据是可以利用来获得竞争优势的资产。因此，它投资了数据治理计划，以确保其数据的质量、准确性和安全性。

美国运通的数据治理计划旨在创建一种在整个组织内一致且高效的数据管理方法。该计划围绕三个关键要素构建：人员、流程和技术。

人员：为了确保项目成功实施，美国运通为组织内的不同个人和部门分配了特定的角色和职责。数据治理团队负责制定和执行数据政策，而业务所有者则负责确保数据的准确性和相关性。该项目还得到了数据管理员网络的支持，他们负责管理各自业务部门内的数据。

流程：美国运通的数据治理计划包括一套标准流程和程序，以确保整个组织的数据管理实践保持一致。这些流程包括数据分类、元数据管理、数据质量控制和数据沿袭跟踪。该计划还包括数据访问、数据共享和数据保留的程序。

技术：为了支持其数据治理计划，美国运通实施了各种工具和技术，包括数据治理门户、数据质量和分析工具以及数据目录。这些工具提供了组织数据资产的集中视图，并允许数据管理员和企业主有效地监控和管理他们的数据。

美国运通数据治理计划实施的时间跨度长达数年。该公司通过对其数据状况进行全面评估开始了数据治理之旅。这项评估帮助该公司了解了其当前的数据管理实践并确定了需要改进的领域。基于此评估，该公司制定了实施数据治理计划的路线图。

在接下来的几年里，美国运通在整个组织内推行了数据治理计划。这包括创建数据治理框架、定义数据政策、建立数据治理角色和职责以及实施数据质量控制。该公司还投资了培训和教育计划，以确保其员工了解数据治理的重要性并具备有效管理数据的必要技能。

3.沃尔玛是全球最大的零售商之一，在全球拥有11,000多家门店。近年来，该公司一直专注于数字化转型，并利用数据来改善客户体验和运营效率。为了实现这一目标，沃尔玛必须实施全面的数据治理计划，以确保整个组织的数据准确、一致且可访问。

沃尔玛面临的主要挑战之一是管理和集成来自多个来源的数据，包括销售点系统、客户数据、供应链数据等。为了克服这一挑战，沃尔玛向领先的数据集成和管理解决方案提供商Informatica寻求帮助。

借助Informatica的工具，沃尔玛能够实施统一的数据集成平台，提供整个组织所有数据的单一视图。该平台使沃尔玛能够简化数据集成、减少数据重复并提高数据质量。此外，该平台还提供高级分析功能，使沃尔玛能够从其数据中获得有价值的见解。

为了确保数据治理计划的成功，沃尔玛还实施了一系列政策和程序，包括数据分类、数据访问控制、数据质量检查和数据保留政策。这些政策由一个集中的数据治理委员会执行，该委员会负责监督数据政策的实施并确保整个组织的合规性。

沃尔玛数据治理计划的成果令人印象深刻。该公司已经能够提高其供应链效率、降低成本并提供更好的客户体验。例如，沃尔玛使用数据分析来优化其库存管理并确保在客户需要时有产品可用。此外，沃尔玛还实施了一系列个性化营销计划，利用客户数据来提供更相关、更有针对性的促销活动。

数据治理对于大型组织至关重要，因为它可以确保数据的有效管理和利用。它使组织能够保持合规性、提高透明度并最大限度地降低与数据使用相关的风险。通过实施有效的数据政策，组织可以创造一种责任感和信任的文化，改善决策，并在其行业中获得竞争优势。

为了实施有效的数据政策，组织必须评估其当前政策，制定全面的计划，明确角色和职责，确保利益相关方的支持，并实施强大的数据质量和完整性控制。定期监测和评估数据政策并为员工提供持续的培训和教育也是成功的关键。

通过案例研究，我们发现成功实施数据治理需要强有力的领导力、跨职能协作以及持续改进的承诺。通过遵循最佳实践并学习成功案例，组织可以制定有效的数据治理政策来支持其数字化转型目标。

投资数据治理不仅是一项合规性要求，也是一项战略决策，可以使组织更加以数据为导向、更具创新性和竞争力。

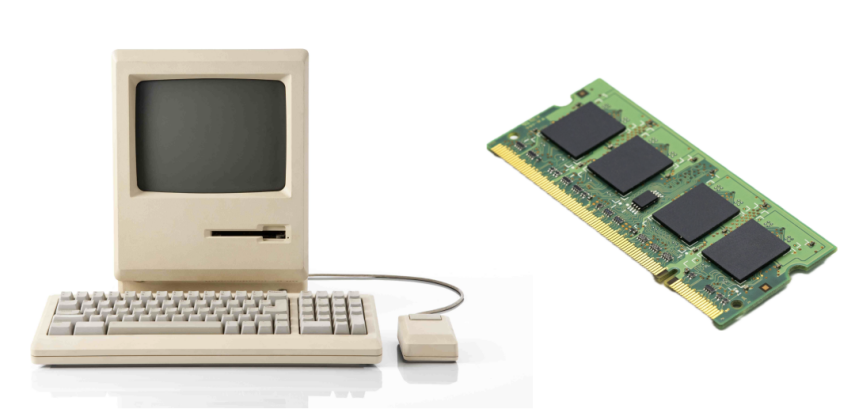
为了实施有效的数据政策，组织需要评估其当前政策，制定全面的计划，明确角色和职责，确保利益相关者和领导层的支持，实施强大的数据质量和完整性控制，定期监控和评估数据政策的有效性，并为员工提供持续的培训和教育。

此外，成功实施需要使用适当的工具和技术来自动化和简化数据治理流程。例如，Informatica等工具使组织能够更有效地管理数据质量、数据集成和元数据管理。

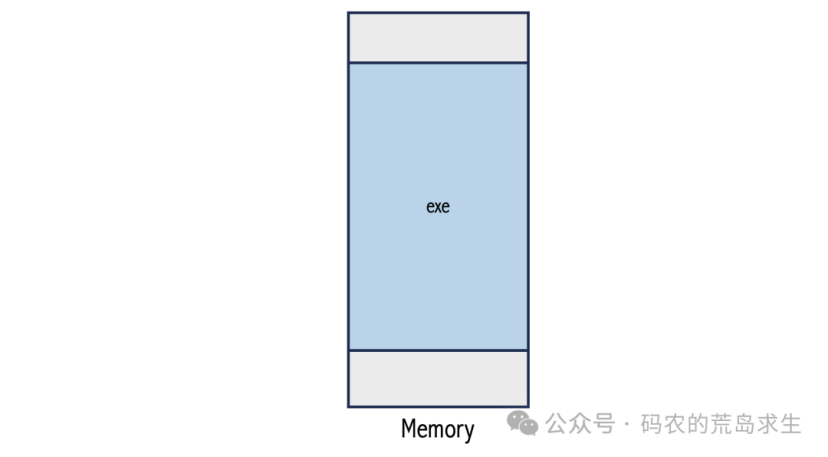
实施有效数据治理政策的组织可以利用其数据资产做出明智决策、优化运营并增强客户体验，从而获得竞争优势。通过遵循本博客中概述的策略和最佳实践，组织可以为成功实施数据治理奠定基础，并在数字时代推动业务增长。

# 内容来源：[动态库和静态库有什么区别？](https://mp.weixin.qq.com/s/9pavORd5qjqEaKN7G_NBmw)

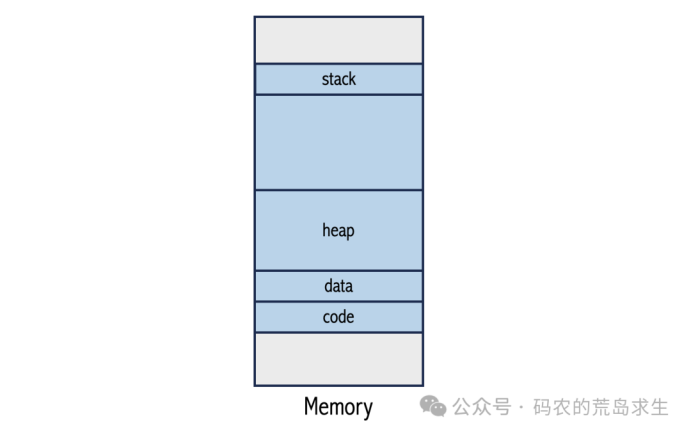
计算机的运行当然离不开内存。



程序运行在内存当中，那么程序在内存中的布局是什么样子的呢？



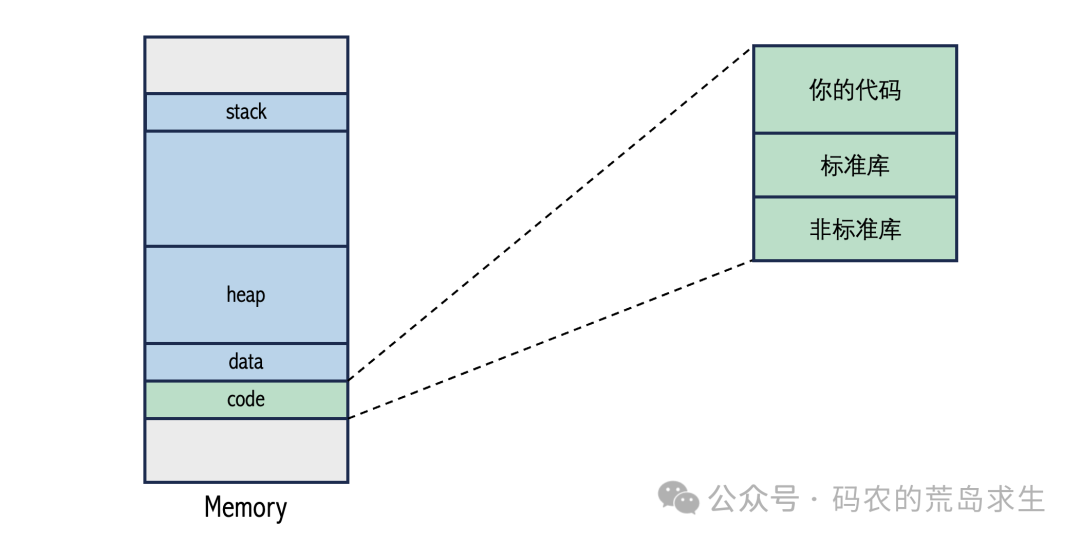
程序的内存分为代码区、数据区、堆区和栈区，它们的布局是这样的，这里重点看代码区。



代码区中是什么呢？

这里主要就是你写的代码，当然还有你使用的库。

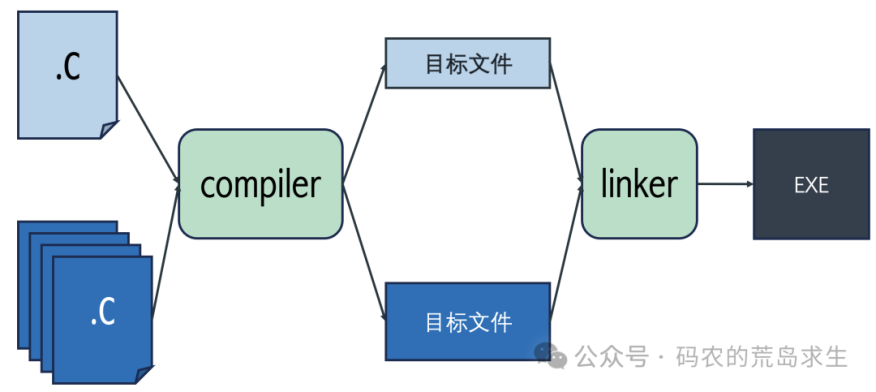
这里主要是标准库，以及非标准库，也就是普通的库。



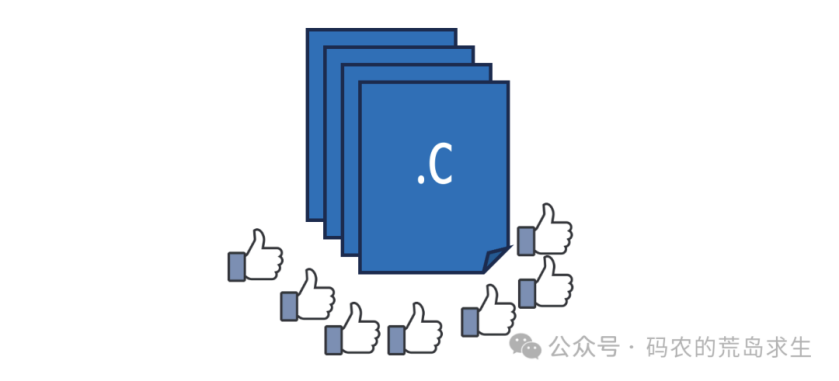
接下来我们来看第一个阶段：编译。

假设你在编写一段空前绝后的代码，需要使用加法函数，考虑到这个函数可能其它人也需要，因此单独放到一个源文件中，其它类似的函数也一样。

接下来是编译过程，先编译成目标文件，链接器把目标文件打包成最终的可执行程序。



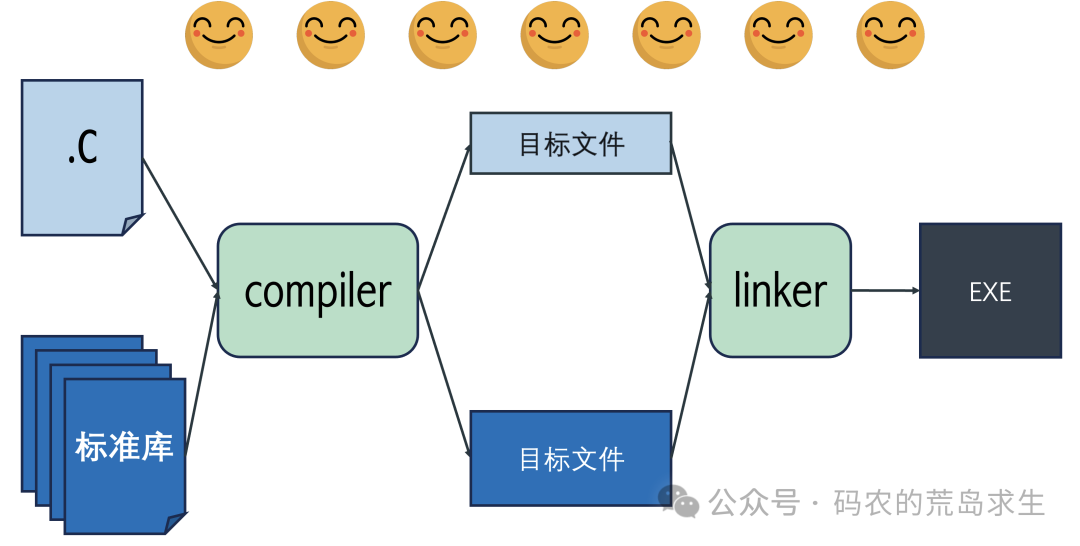
由于你实现的加减乘除等函数非常好用，广受其它程序员的欢迎，因此你写的代码渐渐的变成了一种标准，所有程序员都遵循，标准库就是这么诞生的。



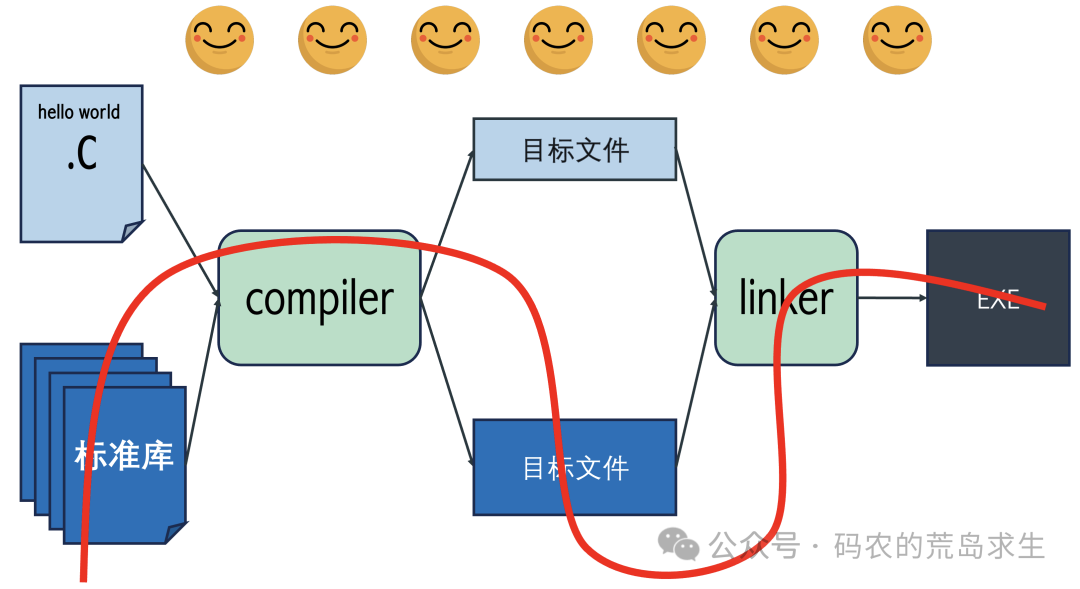
当然，你写的代码也可能根本没有其它人用，只能当做你自己使用的普通库，实际上这才是人生常态，这样的非标准库还是留着自己用吧。

不管是标准库还是非标准库对后续的讲解都是一样的，因此我们以标准库为例来讲解。

现在有了标准库，所有人都很开心，终于不用重复造轮子了。



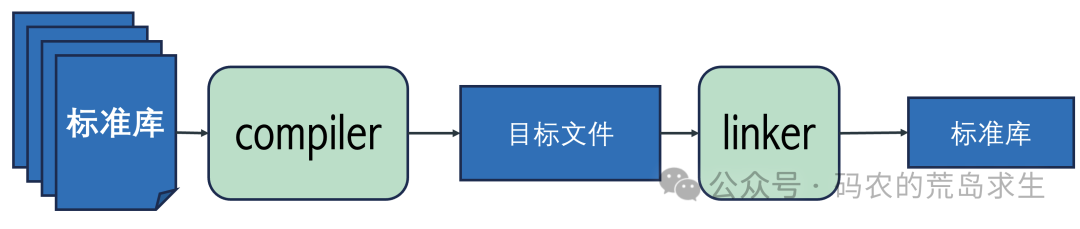
但是很快你会发现，及时写一个最简单的hello world程序，也不得不重新编译一堆标准库代码。



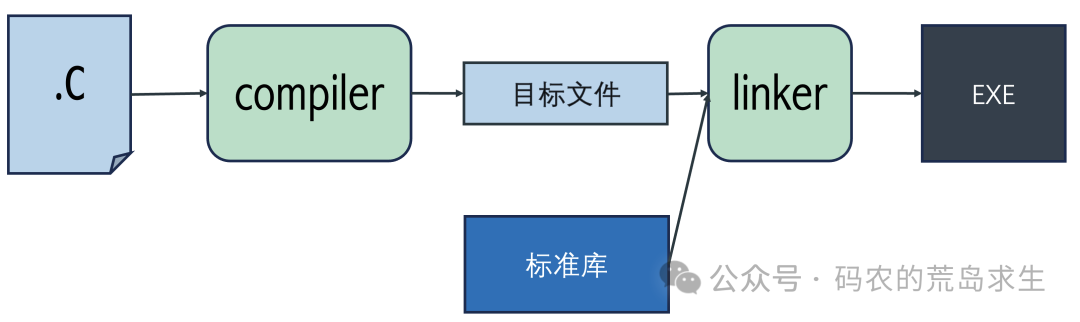
这显然是很浪费时间的，但这对打工人来说是很不错，可以利用编译时间合理摸鱼。

该怎么解决重复编译标准库的问题呢。

很简单，可以提前把标准库编译出来。

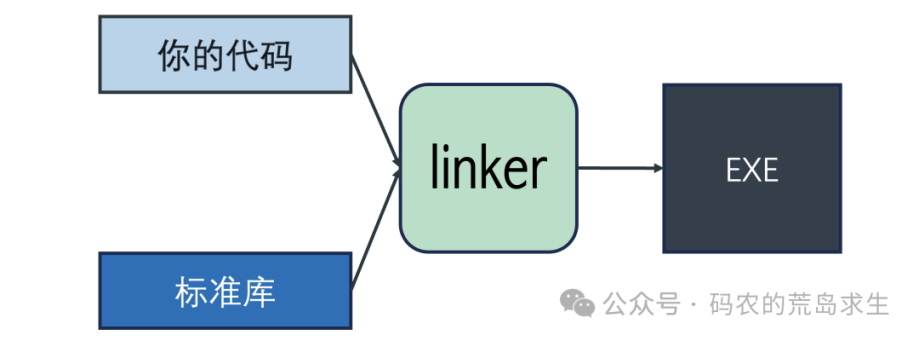


这样我们再次编译自己的项目时直接跳过标准库的编译过程，直接把标准库链接到最终的可执行程序中。

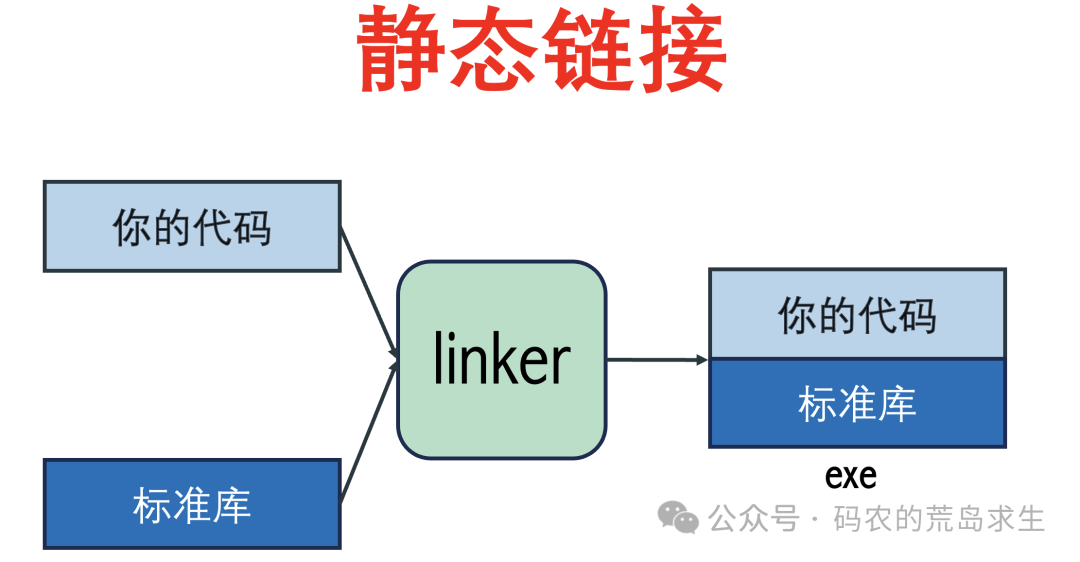


接下来我们来到了第二阶段，链接。

链接器从表面看实际上是一个打包器，把各种目标文件，也就是你的代码和标准库打包起来，如果还有其它库也是一样的道理，经过链接会生成最终的可执行程序。

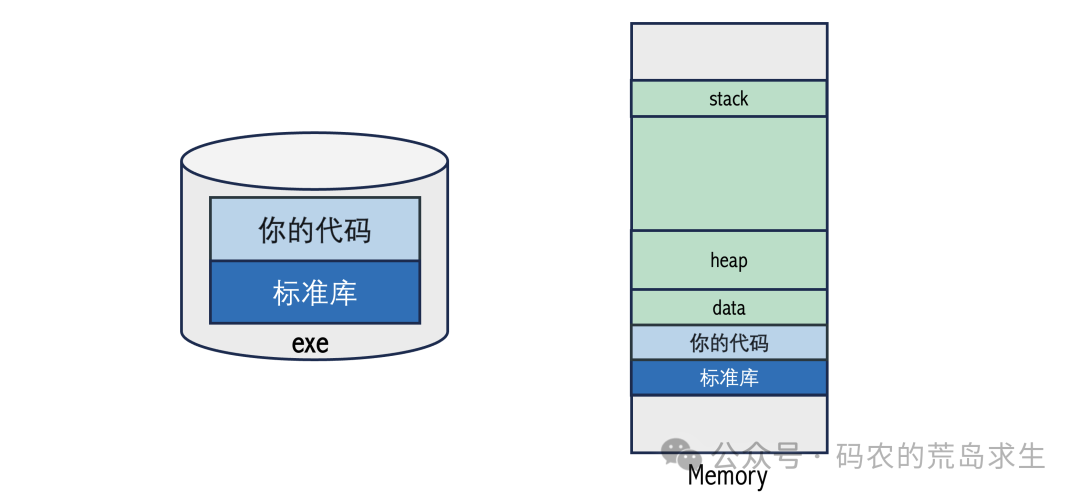


注意看最终的可执行程序中，包含了你的代码和使用的库，这种将所有代码打包到可执行程序的链接方式叫做静态链接。



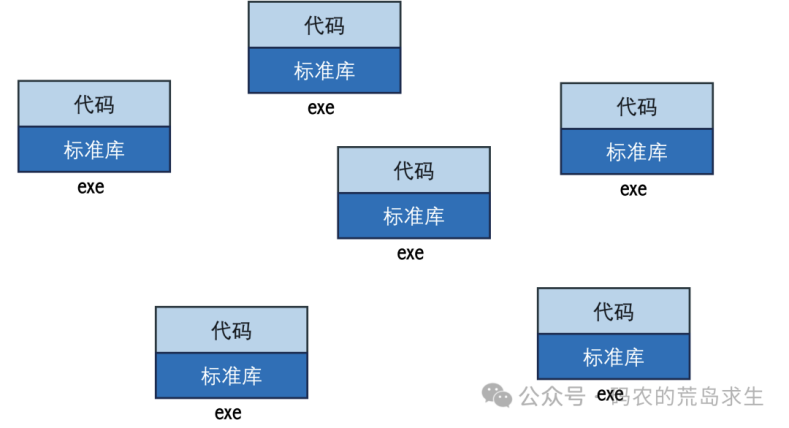
可执行程序当然是保存在磁盘中的，接下来我们双击运行这个程序。

此时会在内存中划分出一块区域，这里就是我们最开始看到的堆区、栈区、数据和代码区。



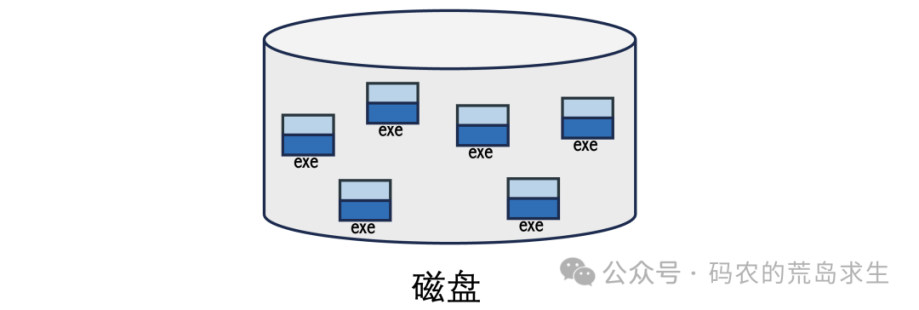
这时会把可执行程序中的指令都拷贝到代码区，这就是程序运行的最基本原理。

显然计算机系统里不会只有一个程序。



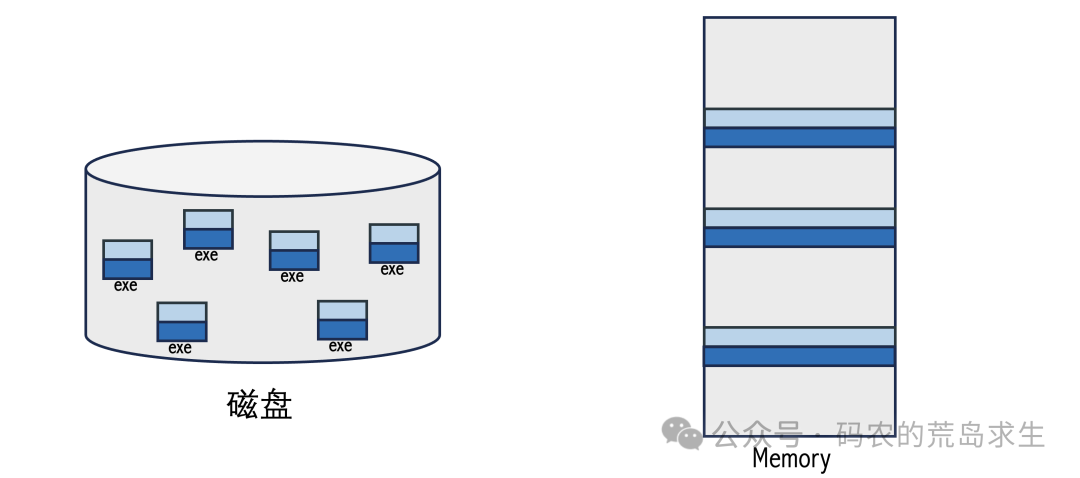
可能会有很多个，如果所有程序都采用静态链接会怎样呢？

如果都是静态链接，那么这些程序都保存一份标准库，由于是标准库，因此它们的内容都是一样的，最终这些程序都保存在磁盘上，占据磁盘空间：



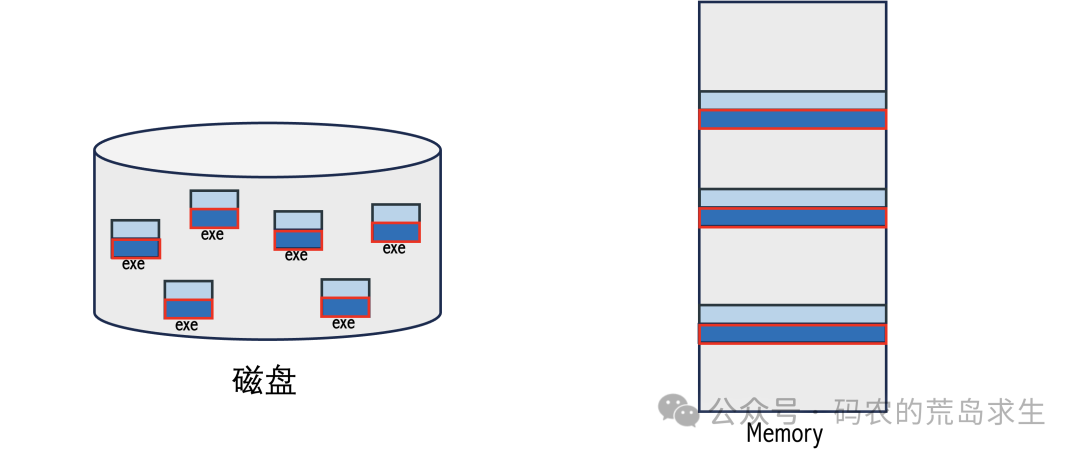
而在运行起来后同样会占用内存空间。

注意看这里和这里，它们的内容都是一样的，这显然是在浪费宝贵的存储资源。



问题出在哪里呢？

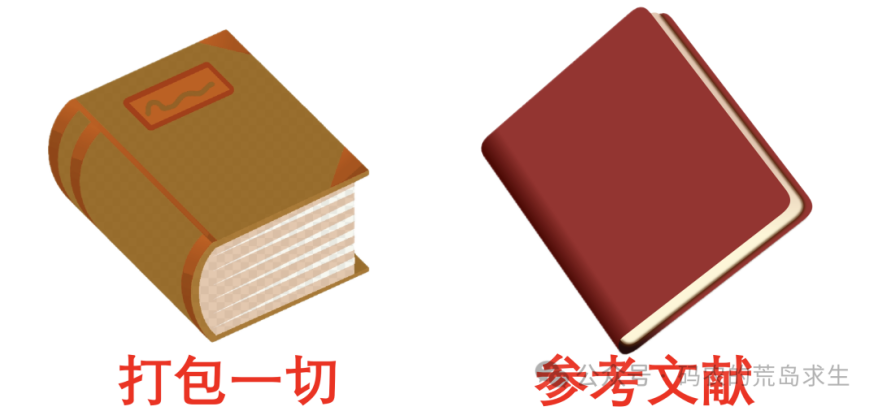
注意看这里，每个可执行程序都包含有一份标准库。



这就像什么呢？这就像一本非常厚的书，这本书把引用到的所有内容到装订在一起。



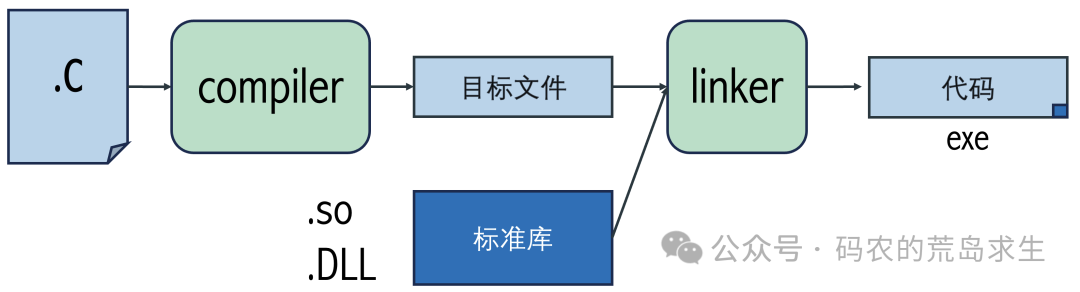
好处当然是更方便读者阅读引用到的内容，但问题就是这样的书太厚重了，更好的方法是把引用到内容放在一页参考文献中而不要打包在一起，这样一本书就很轻量了。



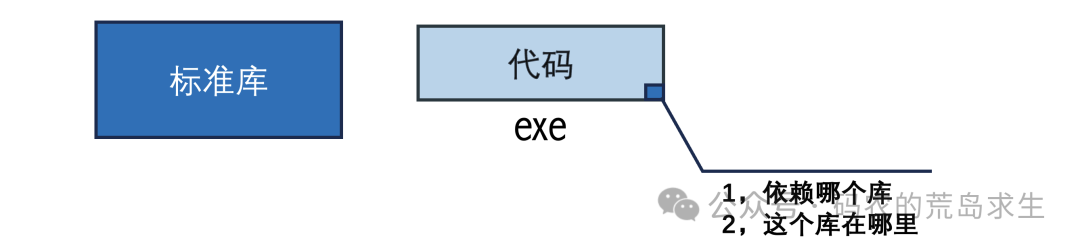
可执行程序也是一样的道理，我们不再打包一切而是使用参考文献。

这就是所谓的动态链接。

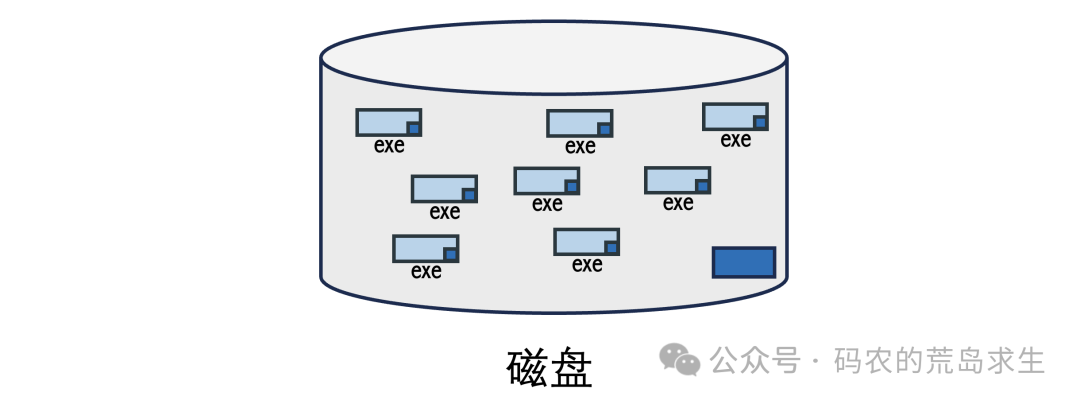
这时提前编译好的库就是动态库，linux下以.so为结尾，windows中是dll文件。



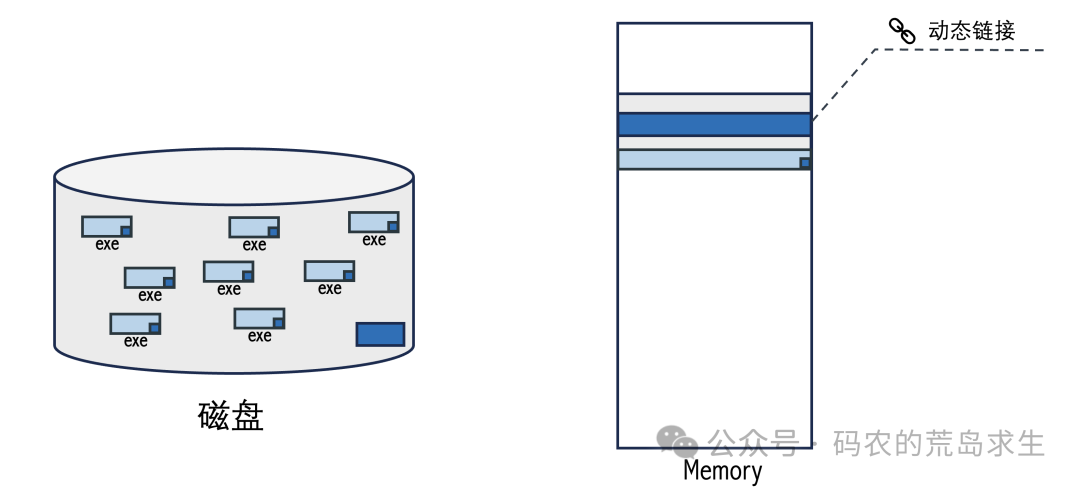
这时的可执行程序中不再包含完整的动态库代码，而是只有一份参考文献，你可以想象一下，这里肯定保存了两项基本信息，依赖了哪个库，这个库保存在了哪里。



这样及时有再多的可执行程序也不怕，因为我们只需要保存一份标准库。



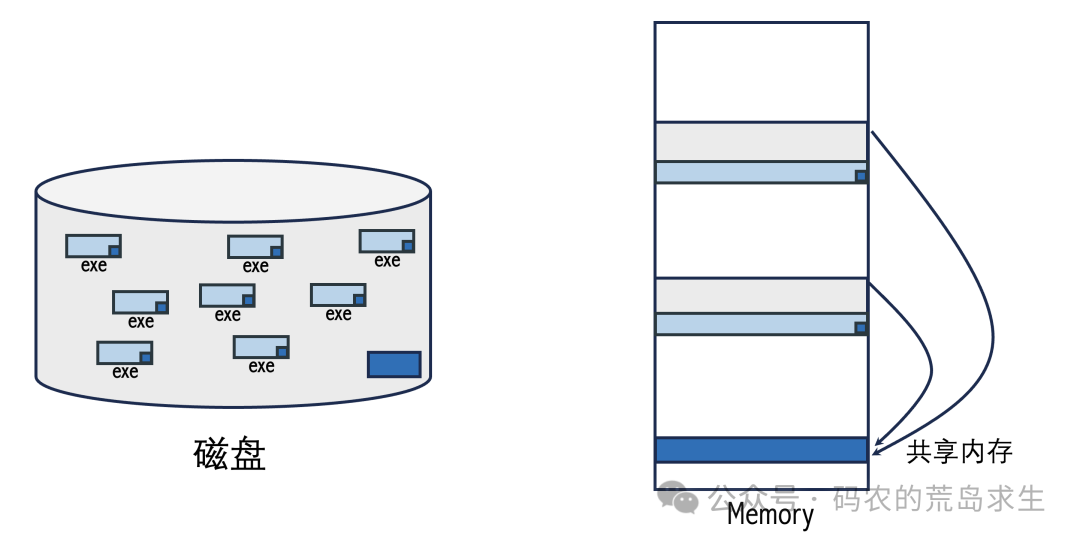
动态链接的程序该怎么运行呢？依然是把可执行程序拷贝到代码区，然后根据参考文献把动态库加载进来，这时才是真正的链接，也就是所谓的动态链接，我们把链接这个过程从编译期推迟到了运行时。



如果是其它程序运行呢？道理也一样。

动态链接下每个程序中不会有一份标准库，而是大家共享内存中的同一份标准库。

这样利用动态链接我们不但节省了磁盘空间也节省了内存空间。



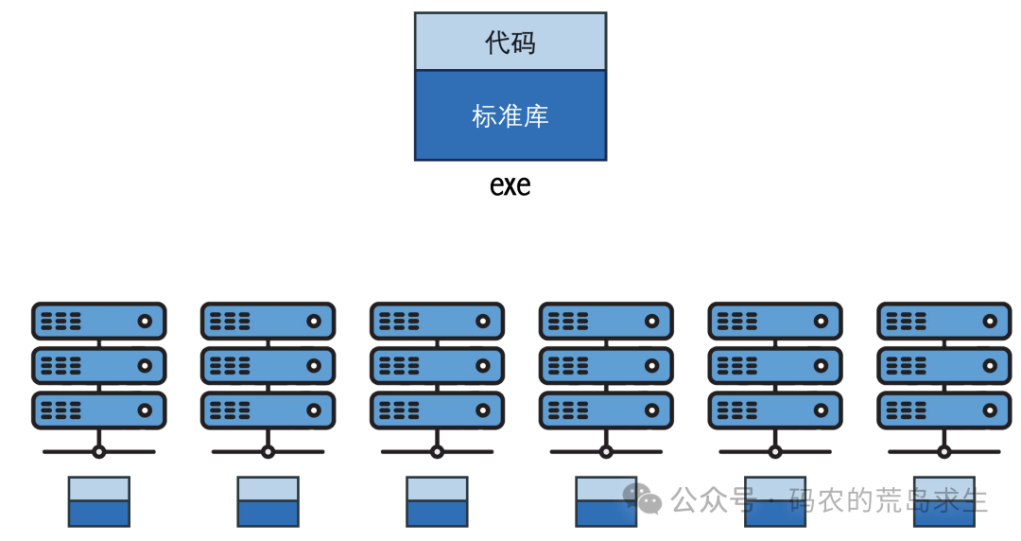
接下来我们简单总结一下静态库和动态库的区别了。

首先是静态链接。

静态链接下的可执行程序中打包了所以用到的东西，就和这个工具箱一样，这样的程序不再依赖其它库。

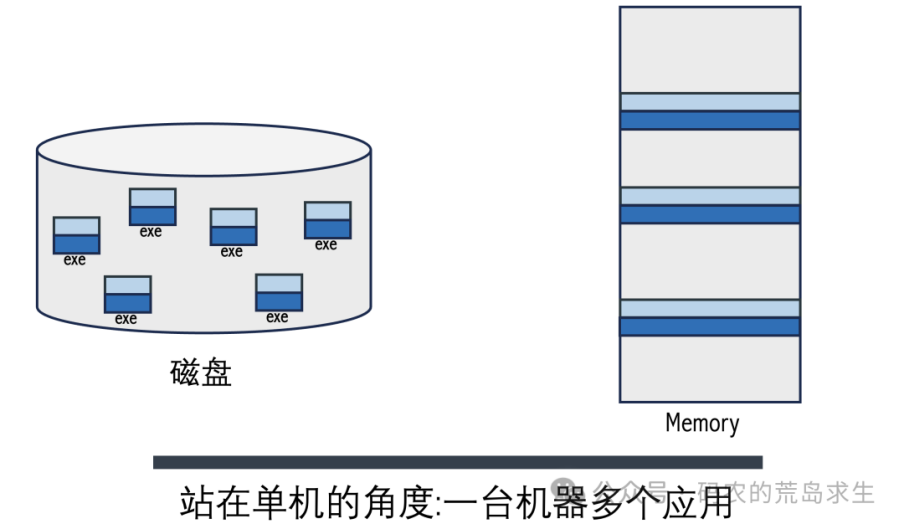


因此在大规模部署到服务器时非常方便，所以你会看到一些互联网服务程序很多采用静态链接。



但是静态链接也有自己缺点，那就是修复bug不是很方便，你需要从头开始编译链接生成最终的可执行程序，然后重新部署服务并重启。

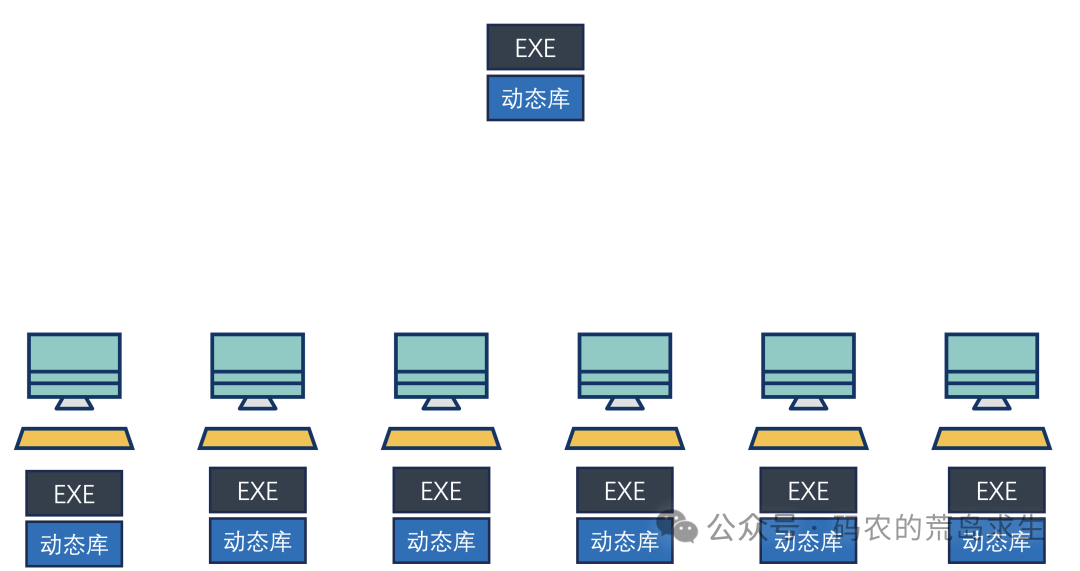
同时静态链接也有我们刚提到的重复占据磁盘和内存空间的问题，而动态库则没有这个问题。



当然我们说静态库的这一问题这是站在单机一台机器有多个应用的角度，而对于一个应用需要部署在多机的角度这些都不是事。

接下来我们看动态库。

假设你用动态链接的方式构建了一个非常受欢迎的应用程序，然后发布出去。



发布出去后发现动态库模块有bug，那么该怎么修复呢？很简单，只需要修复bug后重新生成新的动态库，简单的更新动态库即可，而不需要重新发布应用。



这就是为什么很多应用程序采用动态链接的原因。

# 内容来源：十四届全国人大常委会专题讲座第十讲讲稿《人工智能与智能计算的发展》-孙凝晖

委员长、各位副委员长、秘书长、各位委员：

人工智能领域近年来正在迎来一场由生成式人工智能大模型引领的爆发式发展。2022年11月30日，OpenAI公司推出一款人工智能对话聊天机器人ChatGPT，其出色的自然语言生成能力引起了全世界范围的广泛关注，2个月突破1亿用户，国内外随即掀起了一场大模型浪潮，Gemini、文心一言、Copilot、LLaMA、SAM、SORA等各种大模型如雨后春笋般涌现，2022年也被誉为大模型元年。当前信息时代正加快进入智能计算的发展阶段，人工智能技术上的突破层出不穷，逐渐深入地赋能千行百业，推动人工智能与数据要素成为新质生产力的典型代表。习近平总书记指出，把新一代人工智能作为推动科技跨越发展、产业优化升级、生产力整体跃升的驱动力量，努力实现高质量发展。党的十八大以来，以习近平同志为核心的党中央高度重视智能经济发展，促进人工智能和实体经济深度融合，为高质量发展注入强劲动力。

一、计算技术发展简介

计算技术的发展历史大致可分为四个阶段，算盘的出现标志着人类进入第一代——机械计算时代，第二代——电子计算的标志是出现电子器件与电子计算机，互联网的出现使我们进入第三代——网络计算，当前人类社会正在进入第四阶段——智能计算。

早期的计算装置是手动辅助计算装置和半自动计算装置，人类计算工具的历史是从公元1200年的中国算盘开始，随后出现了纳皮尔筹（1612年）和滚轮式加法器（1642年），到1672年第一台自动完成四则运算的计算装置——步进计算器诞生了。

机械计算时期已经出现了现代计算机的一些基本概念。查尔斯∙巴贝奇（Charles Babbage）提出了差分机（1822年）与分析机（1834年）的设计构想，支持自动机械计算。这一时期，编程与程序的概念基本形成，编程的概念起源于雅卡尔提花机，通过打孔卡片控制印花图案，最终演变为通过计算指令的形式来存储所有数学计算步骤；人类历史的第一个程序员是诗人拜伦之女艾达（Ada），她为巴贝奇差分机编写了一组求解伯努利数列的计算指令，这套指令也是人类历史上第一套计算机算法程序，它将硬件和软件分离，第一次出现程序的概念。

直到在二十世纪上半叶，出现了布尔代数(数学)、图灵机(计算模型)、冯诺依曼体系结构(架构)、晶体管(器件)这四个现代计算技术的科学基础。其中，布尔代数用来描述程序和硬件如CPU的底层逻辑；图灵机是一种通用的计算模型，将复杂任务转化为自动计算、不需人工干预的自动化过程；冯诺依曼体系结构提出了构造计算机的三个基本原则：采用二进制逻辑、程序存储执行、以及计算机由运算器、控制器、存储器、输入设备、输出设备这五个基本单元组成；晶体管是构成基本的逻辑电路和存储电路的半导体器件，是建造现代计算机之塔的“砖块”。基于以上科学基础，计算技术得以高速发展，形成规模庞大的产业。

从1946年世界上第一台电子计算机ENIAC诞生到二十一世纪的今天，已经形成了五类成功的平台型计算系统。当前各领域各种类型的应用，都可以由这五类平台型计算装置支撑。第一类是高性能计算平台，解决了国家核心部门的科学与工程计算问题；第二类是企业计算平台，又称服务器，用于企业级的数据管理、事务处理，当前像百度、阿里和腾讯这些互联网公司的计算平台都属于这一类；第三类是个人电脑平台，以桌面应用的形式出现，人们通过桌面应用与个人电脑交互；第四类是智能手机，主要特点是移动便携，手机通过网络连接数据中心，以互联网应用为主，它们分布式地部署在数据中心和手机终端；第五类是嵌入式计算机，嵌入到工业装备和军事设备，通过实时的控制，保障在确定时间内完成特定任务。这五类装置几乎覆盖了我们信息社会的方方面面，长期以来人们追求的以智能计算应用为中心的第六类平台型计算系统尚未形成。

现代计算技术的发展大致可以划分为三个时代。IT1.0又称电子计算时代（1950-1970），基本特征是以“机”为中心。计算技术的基本架构形成，随着集成电路工艺的进步，基本计算单元的尺度快速微缩，晶体管密度、计算性能和可靠性不断提升，计算机在科学工程计算、企业数据处理中得到了广泛应用。

IT2.0又称网络计算时代（1980-2020），以“人”为中心。互联网将人使用的终端与后台的数据中心连接，互联网应用通过智能终端与人进行交互。以亚马逊等为代表的互联网公司提出了云计算的思想，将后台的算力封装成一个公共服务租借给第三方用户，形成了云计算与大数据产业。

IT3.0又称智能计算时代，始于2020年，与IT2.0相比增加了“物”的概念，即物理世界的各种端侧设备，被数字化、网络化和智能化，实现“人-机-物”三元融合。智能计算时代，除了互联网以外，还有数据基础设施，支撑各类终端通过端边云实现万物互联，终端、物端、边缘、云都嵌入AI，提供与ChatGPT类似的大模型智能服务，最终实现有计算的地方就有AI智能。智能计算带来了巨量的数据、人工智能算法的突破和对算力的爆发性需求。

二、智能计算发展简介

智能计算包括人工智能技术与它的计算载体，大致历经了四个阶段，分别为通用计算装置、逻辑推理专家系统、深度学习计算系统、大模型计算系统。

智能计算的起点是通用自动计算装置（1946年）。艾伦·图灵（Alan Turing）和冯·诺依曼（John von Neumann）等科学家，一开始都希望能够模拟人脑处理知识的过程，发明像人脑一样思考的机器，虽未能实现，但却解决了计算的自动化问题。通用自动计算装置的出现，也推动了1956年人工智能（AI）概念的诞生，此后所有人工智能技术的发展都是建立在新一代计算设备与更强的计算能力之上的。

智能计算发展的第二阶段是逻辑推理专家系统（1990年）。E.A.费根鲍姆（Edward Albert Feigenbaum）等符号智能学派的科学家以逻辑和推理能力自动化为主要目标，提出了能够将知识符号进行逻辑推理的专家系统。人的先验知识以知识符号的形式进入计算机，使计算机能够在特定领域辅助人类进行一定的逻辑判断和决策，但专家系统严重依赖于手工生成的知识库或规则库。这类专家系统的典型代表是日本的五代机和我国863计划支持的306智能计算机主题，日本在逻辑专家系统中采取专用计算平台和Prolog这样的知识推理语言完成应用级推理任务；我国采取了与日本不同的技术路线，以通用计算平台为基础，将智能任务变成人工智能算法，将硬件和系统软件都接入通用计算平台，并催生了曙光、汉王、科大讯飞等一批骨干企业。

符号计算系统的局限性在于其爆炸的计算时空复杂度，即符号计算系统只能解决线性增长问题，对于高维复杂空间问题是无法求解的，从而限制了能够处理问题的大小。同时因为符号计算系统是基于知识规则建立的，我们又无法对所有的常识用穷举法来进行枚举，它的应用范围就受到了很大的限制。随着第二次AI寒冬的到来，第一代智能计算机逐渐退出历史舞台。

直到2014年左右，智能计算进阶到第三阶段——深度学习计算系统。以杰弗里·辛顿（Geoffrey Hinton）等为代表的连接智能学派，以学习能力自动化为目标，发明了深度学习等新AI算法。通过深度神经元网络的自动学习，大幅提升了模型统计归纳的能力，在模式识别等应用效果上取得了巨大突破，某些场景的识别精度甚至超越了人类。以人脸识别为例，整个神经网络的训练过程相当于一个网络参数调整的过程，将大量的经过标注的人脸图片数据输入神经网络，然后进行网络间参数调整，让神经网络输出的结果的概率无限逼近真实结果。神经网络输出真实情况的概率越大，参数就越大，从而将知识和规则编码到网络参数中，这样只要数据足够多，就可以对各种大量的常识进行学习，通用性得到极大的提升。连接智能的应用更加广泛，包括语音识别、人脸识别、自动驾驶等。在计算载体方面，中国科学院计算技术研究所2013年提出了国际首个深度学习处理器架构，国际知名的硬件厂商英伟达（NVIDIA）持续发布了多款性能领先的通用GPU芯片，都是深度学习计算系统的典型代表。

智能计算发展的第四阶段是大模型计算系统（2020年）。在人工智能大模型技术的推动下，智能计算迈向新的高度。2020年，AI从“小模型+判别式”转向“大模型+生成式”，从传统的人脸识别、目标检测、文本分类，升级到如今的文本生成、3D数字人生成、图像生成、语音生成、视频生成。大语言模型在对话系统领域的一个典型应用是OpenAI公司的ChatGPT，它采用预训练基座大语言模型GPT-3，引入3000亿单词的训练语料，相当于互联网上所有英语文字的总和。其基本原理是：通过给它一个输入，让它预测下一个单词来训练模型，通过大量训练提升预测精确度，最终达到向它询问一个问题，大模型产生一个答案，与人即时对话。在基座大模型的基础上，再给它一些提示词进行有监督的指令微调，通过人类的<指令，回复>对逐渐让模型学会如何与人进行多轮对话；最后，通过人为设计和自动生成的奖励函数来进行强化学习迭代，逐步实现大模型与人类价值观的对齐。

大模型的特点是以“大”取胜，其中有三层含义，（1）参数大，GPT-3就有1700亿个参数；（2）训练数据大，ChatGPT大约用了3000亿个单词，570GB训练数据；（3）算力需求大，GPT-3大约用了上万块V100 GPU进行训练。为满足大模型对智能算力爆炸式增加的需求，国内外都在大规模建设耗资巨大的新型智算中心，英伟达公司也推出了采用256个H100芯片，150TB海量GPU内存等构成的大模型智能计算系统。

大模型的出现带来了三个变革。一是技术上的规模定律（Scaling Law），即很多AI模型的精度在参数规模超过某个阈值后模型能力快速提升，其原因在科学界还不是非常清楚，有很大的争议。AI模型的性能与模型参数规模、数据集大小、算力总量三个变量成“对数线性关系”，因此可以通过增大模型的规模来不断提高模型的性能。目前最前沿的大模型GPT-4参数量已经达到了万亿到十万亿量级，并且仍在不断增长中；二是产业上算力需求爆炸式增长，千亿参数规模大模型的训练通常需要在数千乃至数万GPU卡上训练2-3个月时间，急剧增加的算力需求带动相关算力企业超高速发展，英伟达的市值接近两万亿美元，对于芯片企业以前从来没有发生过；三是社会上冲击劳动力市场，北京大学国家发展研究院与智联招聘联合发布的《AI大模型对我国劳动力市场潜在影响研究》报告指出，受影响最大的20个职业中财会、销售、文书位于前列，需要与人打交道并提供服务的体力劳动型工作，如人力资源、行政、后勤等反而相对更安全。

人工智能的技术前沿将朝着以下四个方向发展。第一个前沿方向为多模态大模型。从人类视角出发，人类智能是天然多模态的，人拥有眼、耳、鼻、舌、身、嘴(语言)，从AI视角出发，视觉，听觉等也都可以建模为token的序列，可采取与大语言模型相同的方法进行学习，并进一步与语言中的语义进行对齐，实现多模态对齐的智能能力。

第二个前沿方向为视频生成大模型。OpenAI于2024年2月15日发布文生视频模型SORA，将视频生成时长从几秒钟大幅提升到一分钟，且在分辨率、画面真实度、时序一致性等方面都有显著提升。SORA的最大意义是它具备了世界模型的基本特征，即人类观察世界并进一步预测世界的能力。世界模型是建立在理解世界的基本物理常识（如，水往低处流等）之上，然后观察并预测下一秒将要发生什么事件。虽然SORA要成为世界模型仍然存在很多问题，但可以认为SORA学会了画面想象力和分钟级未来预测能力，这是世界模型的基础特征。

第三个前沿方向为具身智能。具身智能指有身体并支持与物理世界进行交互的智能体，如机器人、无人车等，通过多模态大模型处理多种传感数据输入，由大模型生成运动指令对智能体进行驱动，替代传统基于规则或者数学公式的运动驱动方式，实现虚拟和现实的深度融合。因此，具有具身智能的机器人，可以聚集人工智能的三大流派：以神经网络为代表的连接主义，以知识工程为代表的符号主义和控制论相关的行为主义，三大流派可以同时作用在一个智能体，这预期会带来新的技术突破。

第四个前沿方向是AI4R(AI for Research)成为科学发现与技术发明的主要范式。当前科学发现主要依赖于实验和人脑智慧，由人类进行大胆猜想、小心求证，信息技术无论是计算和数据，都只是起到一些辅助和验证的作用。相较于人类，人工智能在记忆力、高维复杂、全视野、推理深度、猜想等方面具有较大优势，是否能以AI为主进行一些科学发现和技术发明，大幅提升人类科学发现的效率，比如主动发现物理学规律、预测蛋白质结构、设计高性能芯片、高效合成新药等。因为人工智能大模型具有全量数据，具备上帝视角，通过深度学习的能力，可以比人向前看更多步数，如能实现从推断(inference)到推理(reasoning)的跃升，人工智能模型就有潜力具备爱因斯坦一样的想象力和科学猜想能力，极大提升人类科学发现的效率，打破人类的认知边界。这才是真正的颠覆所在。

最后，通用人工智能（Artificial General Intelligence，简称AGI）是一个极具挑战的话题，极具争论性。曾经有一个哲学家和一个神经科学家打赌：25年后（即2023年）科研人员是否能够揭示大脑如何实现意识？当时关于意识有两个流派，一个叫集成信息理论，一个叫全局网络工作空间理论，前者认为意识是由大脑中特定类型神经元连接形成的“结构”，后者指出意识是当信息通过互连网络传播到大脑区域时产生的。2023年，人们通过六个独立实验室进行了对抗性实验，结果与两种理论均不完全匹配，哲学家赢了，神经科学家输了。通过这一场赌约，可以看出人们总是希望人工智能能够了解人类的认知和大脑的奥秘。从物理学的视角看，物理学是对宏观世界有了透彻理解后，从量子物理起步开启了对微观世界的理解。智能世界与物理世界一样，都是具有巨大复杂度的研究对象，AI大模型仍然是通过数据驱动等研究宏观世界的方法，提高机器的智能水平，对智能宏观世界理解并不够，直接到神经系统微观世界寻找答案是困难的。人工智能自诞生以来，一直承载着人类关于智能与意识的种种梦想与幻想，也激励着人们不断探索。

三、人工智能的安全风险

人工智能的发展促进了当今世界科技进步的同时，也带来了很多安全风险，要从技术与法规两方面加以应对。

首先是互联网虚假信息泛滥。这里列举若干场景：一是数字分身。AI Yoon是首个使用DeepFake技术合成的官方“候选人”，这个数字人以韩国国民力量党候选人尹锡悦（Yoon Suk-yeol）为原型，借助尹锡悦20小时的音频和视频片段、以及其专门为研究人员录制的3000多个句子，由当地一家DeepFake技术公司创建了虚拟形象AI Yoon，并在网络上迅速走红。实际上AI Yoon表达的内容是由竞选团队撰写的，而不是候选人本人。

二是伪造视频，尤其是伪造领导人视频引起国际争端，扰乱选举秩序，或引起突发舆情事件，如伪造尼克松宣布第一次登月失败，伪造乌克兰总统泽连斯基宣布“投降”的信息，这些行为导致新闻媒体行业的社会信任衰退。

三是伪造新闻，主要通过虚假新闻自动生成牟取非法利益，使用ChatGPT生成热点新闻，赚取流量，截至2023年6月30日全球生成伪造新闻网站已达277个，严重扰乱社会秩序。

四是换脸变声，用于诈骗。如由于AI语音模仿了企业高管的声音，一家香港国际企业因此被骗3500万美元。

五是生成不雅图片，特别是针对公众人物。如影视明星的色情视频制作，造成不良社会影响。因此，迫切需要发展互联网虚假信息的伪造检测技术。

其次，AI大模型面临严重可信问题。这些问题包括：（1）“一本正经胡说八道”的事实性错误；（2）以西方价值观叙事，输出政治偏见和错误言论；（3）易被诱导，输出错误知识和有害内容；（4）数据安全问题加重，大模型成为重要敏感数据的诱捕器，ChatGPT将用户输入纳入训练数据库，用于改善ChatGPT，美方能够利用大模型获得公开渠道覆盖不到的中文语料，掌握我们自己都可能不掌握的“中国知识”。因此，迫切需要发展大模型安全监管技术与自己的可信大模型。

除了技术手段外，人工智能安全保障需要相关立法工作。2021年科技部发布《新一代人工智能伦理规范》，2022年8月，全国信息安全标准化技术委员会发布《信息安全技术 机器学习算法安全评估规范》，2022-2023年，中央网信办先后发布《互联网信息服务算法推荐管理规定》《互联网信息服务深度合成管理规定》《生成式人工智能服务管理办法》等。欧美国家也先后出台法规，2018年5月25日，欧盟出台《通用数据保护条例》，2022年10月4日，美国发布《人工智能权利法案蓝图》，2024年3月13日，欧洲议会通过了欧盟《人工智能法案》。

我国应加快推进《人工智能法》出台，构建人工智能治理体系，确保人工智能的发展和应用遵循人类共同价值观，促进人机和谐友好；创造有利于人工智能技术研究、开发、应用的政策环境；建立合理披露机制和审计评估机制，理解人工智能机制原理和决策过程；明确人工智能系统的安全责任和问责机制，可追溯责任主体并补救；推动形成公平合理、开放包容的国际人工智能治理规则。

四、中国智能计算发展困境

人工智能技术与智能计算产业处于中美科技竞争的焦点，我国在过去几年虽然取得了很大的成绩，但依然面临诸多发展困境，特别是由美国的科技打压政策带来的困难。

困境一为美国在AI核心能力上长期处于领先地位，中国处于跟踪模式。中国在AI高端人才数量、AI基础算法创新、AI底座大模型能力（大语言模型、文生图模型、文生视频模型）、底座大模型训练数据、底座大模型训练算力等，都与美国存在一定的差距，并且这种差距还将持续很长一段时间。

困境二为高端算力产品禁售，高端芯片工艺长期被卡。A100，H100，B200等高端智算芯片对华禁售。华为、龙芯、寒武纪、曙光、海光等企业都进入实体清单，它们芯片制造的先进工艺受限，国内可满足规模量产的工艺节点落后国际先进水平2-3代，核心算力芯片的性能落后国际先进水平2-3代。

困境三为国内智能计算生态孱弱，AI开发框架渗透率不足。英伟达CUDA(Compute Unified Device Architecture, 通用计算设备架构)生态完备，已形成了事实上的垄断。国内生态孱弱，具体表现在：一是研发人员不足，英伟达CUDA生态有近2万人开发，是国内所有智能芯片公司人员总和的20倍；二是开发工具不足，CUDA有550个SDK(Software Development Kit, 软件开发工具包)，是国内相关企业的上百倍；三是资金投入不足，英伟达每年投入50亿美元，是国内相关公司的几十倍；四是AI开发框架TensorFlow占据工业类市场，PyTorch占据研究类市场，百度飞桨等国产AI开发框架的开发人员只有国外框架的1/10。更为严重的是国内企业之间山头林立，无法形成合力，从智能应用、开发框架、系统软件、智能芯片，虽然每层都有相关产品，但各层之间没有深度适配，无法形成一个有竞争力的技术体系。

困境四为AI应用于行业时成本、门槛居高不下。当前我国AI应用主要集中在互联网行业和一些国防领域。AI技术推广应用于各行各业时，特别是从互联网行业迁移到非互联网行业，需要进行大量的定制工作，迁移难度大，单次使用成本高。最后，我国在AI领域的人才数量与实际需求相比也明显不足。

五、中国如何发展智能计算的道路选择

人工智能发展的道路选择对我国至关重要，关系到发展的可持续性与最终的国际竞争格局。当前人工智能的使用成本十分高昂，微软Copilot套件要支付每月10美元的使用费用，ChatGPT每天消耗50万千瓦时的电力，英伟达B200芯片价格高达3万美元以上。总体来说，我国应发展用得起、安全可信的人工智能技术，消除我国信息贫困人口、并造福“一带一路”国家；低门槛地赋能各行各业，让我国的优势产业保持竞争力，让相对落后的产业能够大幅地缩小差距。

选择一：统一技术体系走闭源封闭，还是开源开放的道路？

支撑智能计算产业的是一个相互紧耦合的技术体系，即由一系列技术标准和知识产权将材料、器件、工艺、芯片、整机、系统软件、应用软件等密切联系在一起的技术整体。我国发展智能计算技术体系存在三条道路：

一是追赶兼容美国主导的A体系。我国大多数互联网企业走的是GPGPU/CUDA兼容道路，很多芯片领域的创业企业在生态构建上也是尽量与CUDA兼容，这条道路较为现实。由于在算力方面美国对我国工艺和芯片带宽的限制，在算法方面国内生态林立很难形成统一，生态成熟度严重受限，在数据方面中文高质量数据匮乏，这些因素会使得追赶者与领先者的差距很难缩小，一些时候还会进一步拉大。

二是构建专用封闭的B体系。在军事、气象、司法等专用领域构建企业封闭生态，基于国产成熟工艺生产芯片，相对于底座大模型更加关注特定领域垂直类大模型，训练大模型更多采用领域专有高质量数据等。这条道路易于形成完整可控的技术体系与生态，我国一些大型骨干企业走的是这条道路，它的缺点是封闭，无法凝聚国内大多数力量，也很难实现全球化。

三是全球共建开源开放的C体系。用开源打破生态垄断，降低企业拥有核心技术的门槛，让每个企业都能低成本地做自己的芯片，形成智能芯片的汪洋大海，满足无处不在的智能需求。用开放形成统一的技术体系，我国企业与全球化力量联合起来共建基于国际标准的统一智能计算软件栈。形成企业竞争前共享机制，共享高质量数据库，共享开源通用底座大模型。对于全球开源生态，我国企业在互联网时代收益良多，我国更多的是使用者，是参与者，在智能时代我国企业在RISC-V⑥+AI开源技术体系上应更多地成为主力贡献者，成为全球化开放共享的主导力量。

选择二：拼算法模型，还是拼新型基础设施？

人工智能技术要赋能各行各业，具有典型的长尾效应。我国80%的中小微企业，需要的是低门槛、低价格的智能服务。因此，我国智能计算产业必须建立在新的数据空间基础设施之上，其中关键是我国应率先实现智能要素即数据、算力、算法的全面基础设施化。这项工作可比肩二十世纪初美国信息高速公路计划（即信息基础设施建设）对互联网产业的历史作用。

信息社会最核心的生产力是网络空间(Cyberspace)。网络空间的演进过程是：从机器一元连接构成的计算空间，演进到人机信息二元连接构成的信息空间，再演进到人机物数据三元连接构成的数据空间。从数据空间看，人工智能的本质是数据的百炼成钢，大模型就是对互联网全量数据进行深度加工后的产物。在数字化时代，在互联网上传输的是信息流，是算力对数据进行粗加工后的结构化抽象；在智能时代，在互联网上传输的是智能流，是算力对数据进行深度加工与精炼后的模型化抽象。智能计算的一个核心特征就是用数值计算、数据分析、人工智能等算法，在算力池中加工海量数据件，得到智能模型，再嵌入到信息世界、物理世界的各个过程中。

我国政府已经前瞻性地提前布局了新型基础设施，在世界各国竞争中抢占了先机。首先，数据已成为国家战略信息资源。数据具有资源要素与价值加工两重属性，数据的资源要素属性包括生产、获取、传输、汇聚、流通、交易、权属、资产、安全等各个环节，我国应继续加大力度建设国家数据枢纽与数据流通基础设施。

其次，AI大模型就是数据空间的一类算法基础设施。以通用大模型为基座，构建大模型研发与应用的基础设施，支撑广大企业研发领域专用大模型，服务于机器人、无人驾驶、可穿戴设备、智能家居、智能安防等行业，覆盖长尾应用。

最后，全国一体化算力网建设在推动算力的基础设施化上发挥了先导作用。算力基础设施化的中国方案，应在大幅度降低算力使用成本和使用门槛的同时，为最广范围覆盖人群提供高通量、高品质的智能服务。算力基础设施的中国方案需要具备“两低一高”，即在供给侧，大幅度降低算力器件、算力设备、网络连接、数据获取、算法模型调用、电力消耗、运营维护、开发部署的总成本，让广大中小企业都消费得起高品质的算力服务，有积极性开发算力网应用；在消费侧，大幅度降低广大用户的算力使用门槛，面向大众的公共服务必须做到易获取、易使用，像水电一样即开即用，像编写网页一样轻松定制算力服务，开发算力网应用。在服务效率侧，中国的算力服务要实现低熵高通量，其中高通量是指在实现高并发度服务的同时，端到端服务的响应时间可满足率高；低熵是指在高并发负载中出现资源无序竞争的情况下，保障系统通量不急剧下降。保障“算得多”对中国尤其重要。

选择三：AI+着重赋能虚拟经济，还是发力实体经济？

“AI+”的成效是人工智能价值的试金石。次贷危机后，美国制造业增加值占GDP的比重从1950年的28%降低为2021年的11%，美国制造业在全行业就业人数占比从1979年的35%降低为2022年的8%，可见美国更倾向于回报率更高的虚拟经济，轻视投资成本高且经济回报率低的实体经济。中国倾向于实体经济与虚拟经济同步发展，更加重视发展装备制造、新能源汽车、光伏发电、锂电池、高铁、5G等实体经济。

相应地美国AI主要应用于虚拟经济和IT基础工具，AI技术也是“脱实向虚”，自2007年以来硅谷不断炒作虚拟现实（Virtual Reality，VR）、元宇宙、区块链、Web3.0、深度学习、AI大模型等，是这个趋势的反映。

我国的优势在实体经济，制造业全球产业门类最齐全，体系最完整，特点是场景多、私有数据多。我国应精选若干行业加大投入，形成可低门槛全行业推广的范式，如选择装备制造业作为延续优势代表性行业，选择医药业作为快速缩短差距的代表性行业。赋能实体经济的技术难点是AI算法与物理机理的融合。

人工智能技术成功的关键是能否让一个行业或一个产品的成本大幅下降，从而将用户数与产业规模扩大10倍，产生类似于蒸汽机对于纺织业，智能手机对于互联网业的变革效果。

我国应走出适合自己的人工智能赋能实体经济的高质量发展道路。

（主讲人系中国工程院院士，中国科学院计算技术研究所研究员、学术委员会主任）

# 内容来源：[收藏级｜蚂蚁金服EB级大数据治理最佳实践](https://mp.weixin.qq.com/s/HKof4zgnwmaLH1O_aU6i2A)

**导读**

本文将分享蚂蚁集团在大数据治理实践过程中沉淀的经验。

**主要分成四个部分：**

1.数据治理概况

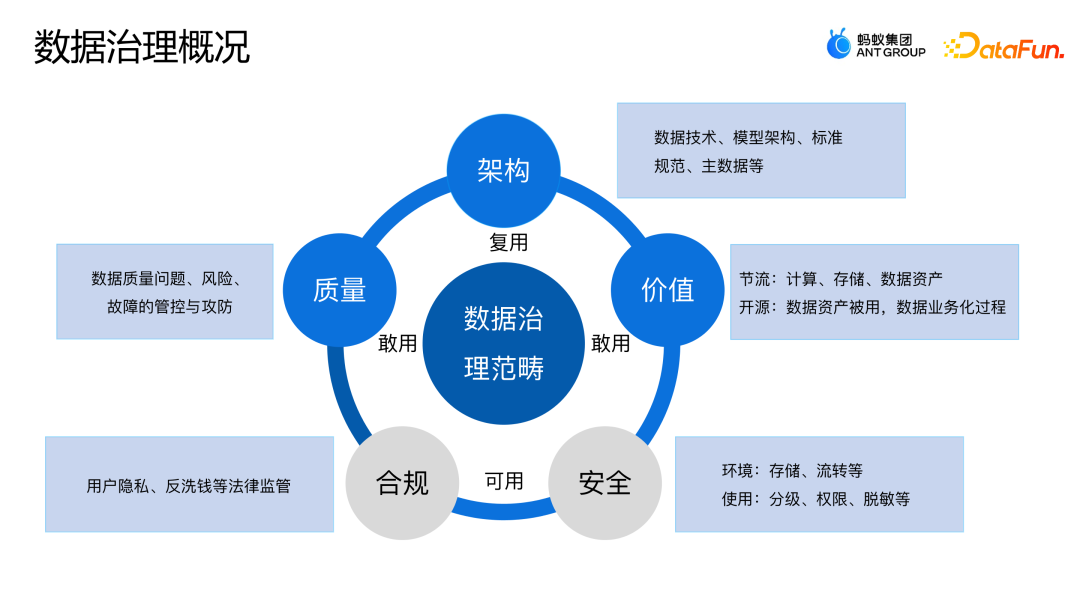
2.数据质量治理

3.数据计存治理

4.对数据治理未来的思考

**01数据治理概况**

业界对于数据治理的定义有很多种，蚂蚁在数据治理时主要关注对企业运转非常关键的架构、安全、合规、质量和价值这五个方面。



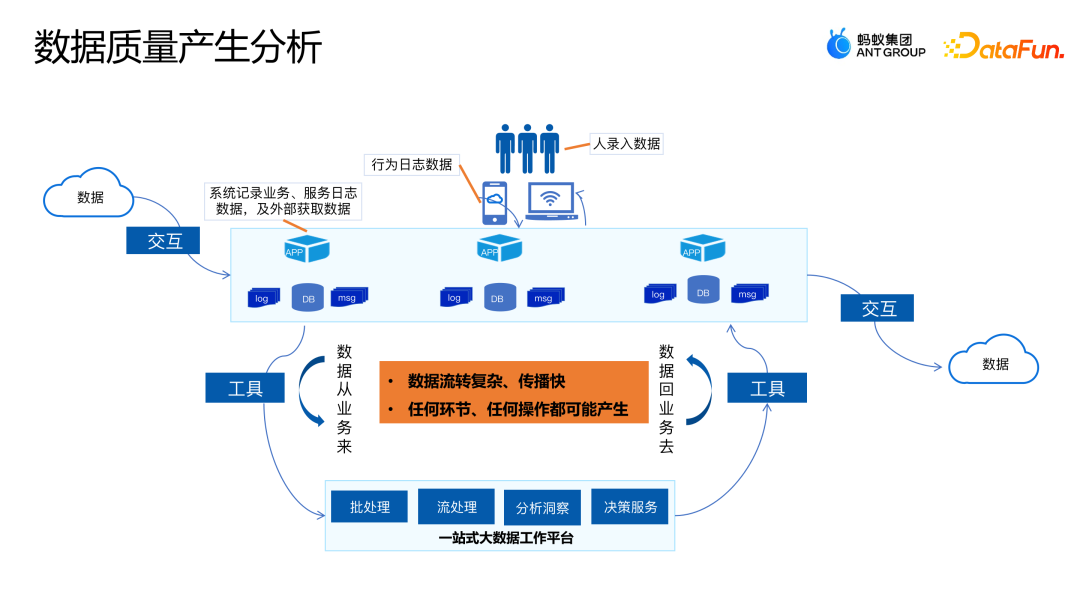
为什么是这五个方面呢？

* 首先，要保证整个数据在业务上是可以流转起来的、是可用的，包含两个基本要求：首先是要符合最近关注度非常高的用户隐私、反洗钱等监管法律的要求，保障数据是合规的；第二是要保证数据在各个环境上的存储、流转和使用都是安全的。这些是在安全合规领域要重点去解决的问题。
* 其次，交付给业务的数据不能错漏，也不能延迟，这属于数据质量范畴，这个领域主要解决让业务敢用数据的问题。
* 另外，大数据领域有非常多的人在协同开发，希望产出的数据是有序的，既是可复用的又是好用的，所以，需要重点做好数据架构的规划和治理，包括数据模型设计、数据标准规范和主数据等。
* 最后，数据是一个闭环的生态，从拿到数据到加工数据，再到赋能业务，希望整个过程是可持续的，在这个可持续的过程中需要有数据价值的体现。价值可以分成两类，一类是负向的价值成本，包括数据运转过程中计算、存储、数据资产带来的机器资源成本；另一类是正向的业务价值，是指数据被使用消费过程中发挥的价值。业界一直在关注数据的正向价值，从数据要素来讲，核心是将数据从原来的资源或者产品，转化成面向未来的商品。对数据价值的衡量是未来一大趋势。

本次分享聚焦于其中的两个命题：数据质量治理和计存治理。接下来将分别进行介绍。

**02数据质量治理**

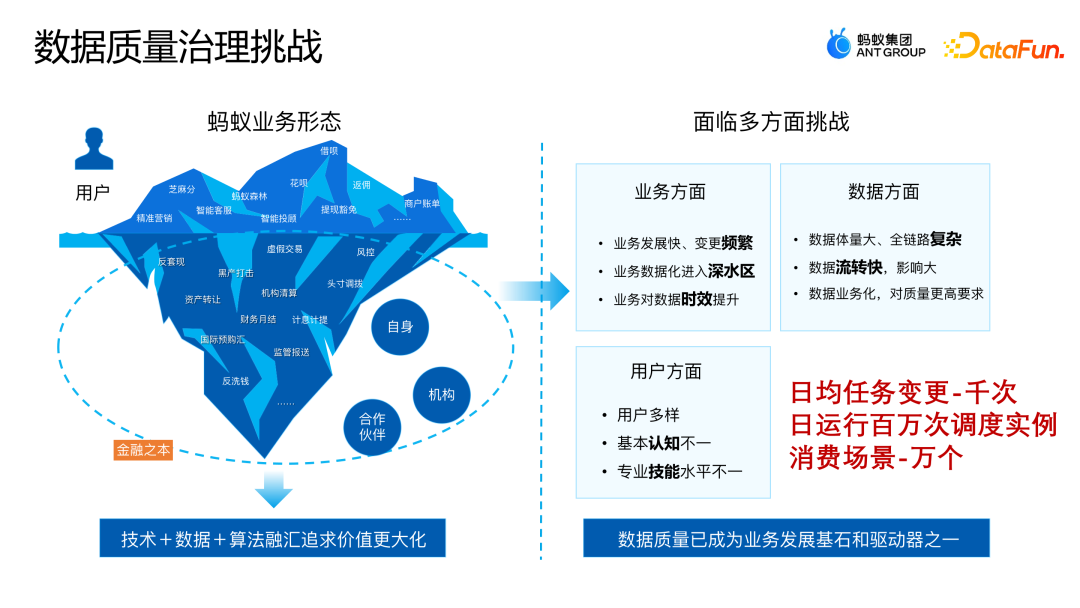
**1.数据质量产生分析**



蚂蚁的数据来源众多，包括行为日志、系统服务端收集的数据等。从类型上看，有DB类、日志类、log类等，还有消息类的和非结构化的数据。大模型出来之后，我们通过一系列工具，将这些数据都存储到了蚂蚁一站式的大数据工作平台上，经过批流的处理进行分析洞察、决策服务。也就是说，数据从业务中来，通过模型算法加工，最终又回到了业务中去。整个流转过程非常复杂，涉及到很多的工具引擎，中间任何环节和操作都可能引发数据质量问题。提供给业务的数据错了、漏了或者延迟了，是经常遇到的一个痛点。

**2.数据质量治理挑战**

在介绍蚂蚁如何进行数据质量治理之前，先来了解一下蚂蚁的业务形态。第一部分是大家感知的“冰山之上”的C端业务，包含芝麻分、蚂蚁森林、花呗、借呗等；第二部分是面向机构监管的“冰山之下”的业务，包括机构清算、计息、计提等，这些业务需要大量的技术支撑，甚至是数据加算法融汇，以追求价值的最大化。在金融业务极度严苛的要求下，做好整体的数据质量保障是非常重要的。



数据质量治理面临着诸多挑战，主要包括：

* 业务方面：蚂蚁业务发展快，变更非常多，任何一次变更出错都可能有很大的影响。无论从用户体验，还是智能化角度，对数据产出的时效都有非常高的要求。
* 数据方面：大部分是金融层面的业务，对数据质量的要求也非常高。
* 用户方面：整条链路上有非常多的角色参与，比如有BI团队、技术团队、数据团队和产品运营团队等等。每个人的基本认知和专业水平都不一样，人为操作可能也会带来一定的风险。

目前蚂蚁整体日均变更任务量在几千次以上，每天日运行任务调度实例达到了百万次以上，数据应用的核心消费场景有数万个，数据质量已经成为蚂蚁业务发展的基石和驱动器之一。这也是为什么今天蚂蚁非常重视数据质量建设的原因。

**3.数据质量顶层设计**

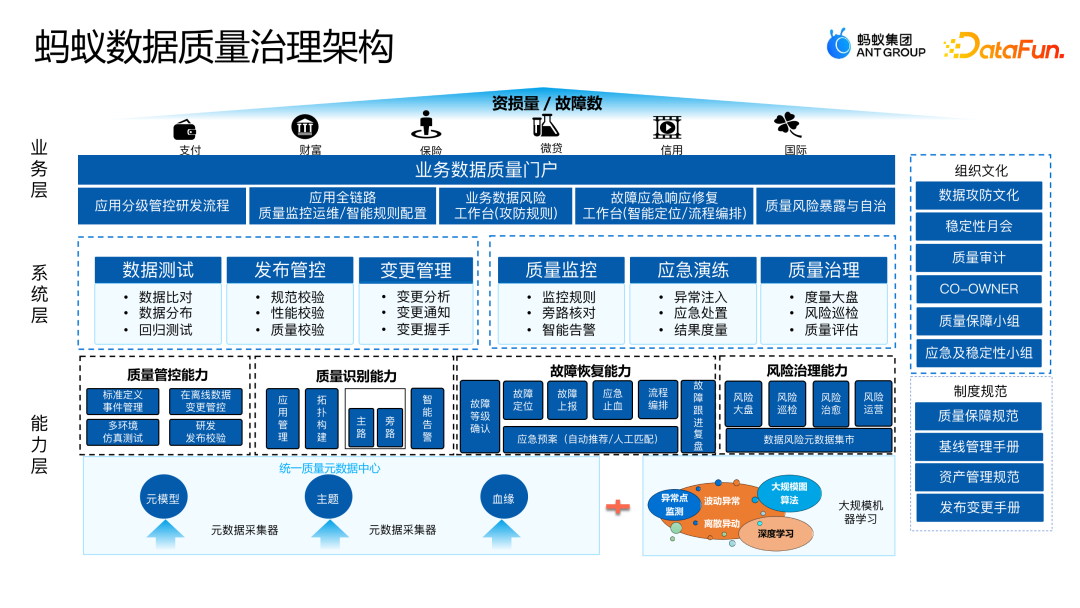
在这么复杂的情况下，怎么解决数据质量的问题呢？单点处理问题很难全面保障数据质量，很有可能拆东墙补西墙，或者这里解决了那里却漏掉了。进行全面的数据质量治理，需要有良好的顶层设计，我们将风险分成三类：数据技术引擎风险、数据内容风险及数据应用风险。



具体落地的核心思路如下。首先，保障目标重点聚焦于高可用和资金安全业务场景：

* 事前，做到整体的研发质量保障，包括测试、仿真等工作；
* 事中，重点解决变更风险的管控；
* 事后，当出现问题的时候，要确保整个生产运行是高可用的，需要重点建设主动发现和快速恢复的能力。
* 另外，还成立了数据和技术的联合蓝军对整个保障体系去做攻击，来验证布防是否可靠。

**4. 数据质量治理架构**



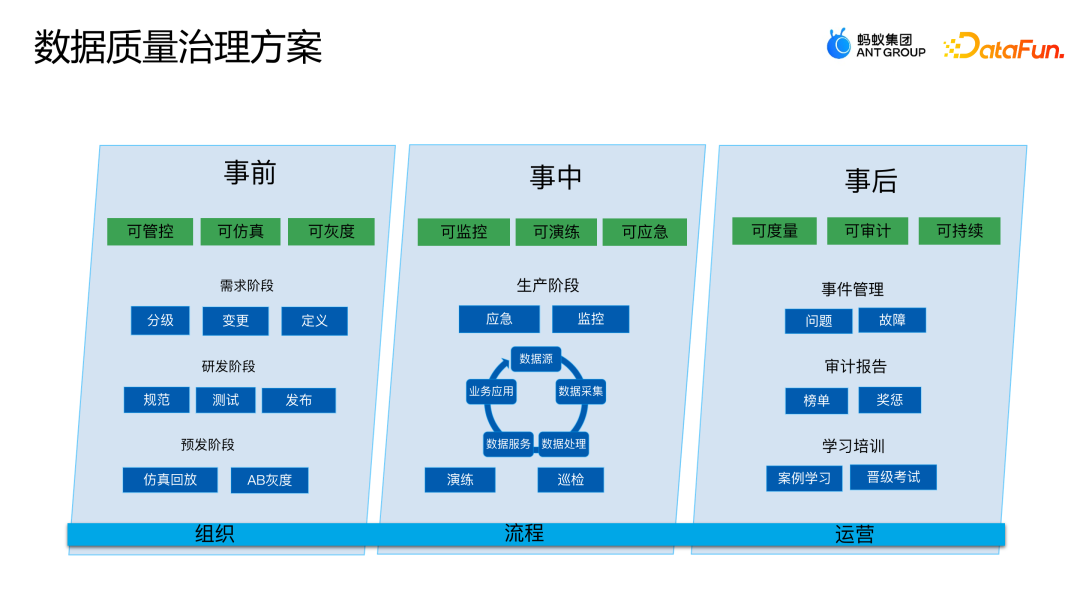
从纵向来看，蚂蚁的数据质量治理架构总体分为三层：

* 能力层，包含质量管控、质量识别、故障恢复和风险治理的能力，并建立了统一质量元数据中心，为后面AI加质量的尝试及相关能力的演进打下了一个非常好的基础。建议在做质量风险保障时，要重视元数据的建设，而且前期就要做好规划。同时，围绕元数据，我们结合大规模机器学习等算法去尝试探索智能化的波动、异常、离散等异常及风险点的识别。
* 系统层，主要围绕数据测试、发布管控、变更管理、质量监控、应急演练和质量治理建设六大产品的能力。
* 业务层，作为数据中台，产品能力开放给业务数据团队、质量团队使用，帮助建设每个业务数据质量的门户，包含整个应用分级管控研发流程、全链路的质量监控运维平台等。

从横向来看，质量治理贯穿全链路系统，并建设了组织文化和制度规范。组织文化包含数据攻防、质量审计、质量保障小组等，做到了全局高效拉通。制度规范包含质量保障规范、基线管理手册、发布变更手册等，形成了全局制度上的规范。在整个实施过程中，重点是以止损量/故障数核心指标为抓手，发现保障体系里面的问题，通过核心指标驱动整个体系持续地迭代和优化。

**5.数据质量治理方案**

接下来深入介绍数据质量治理围绕事前、事中、事后的技术能力。技术上处理离线数据故障有一个核心目标——“五分钟内发现故障，五十分钟内恢复执行”。处理线上数据故障的目标是“一分钟发现问题，五分钟定位问题，十分钟恢复执行”。之所以离线和线上的目标不同，是因为离线数据整条链路比较长，定位和恢复需要较长的时间，另外，当前的故障发现能力、元数据时效性等也存在一定局限性。



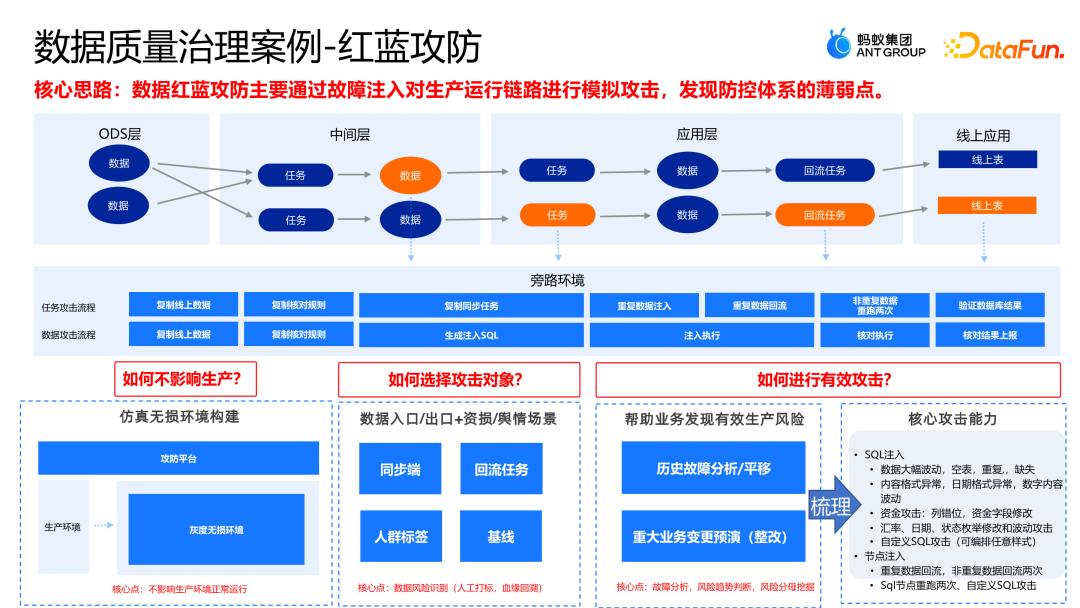
执行的核心策略包括事前、事中、事后三部分。

* 事前要做到可管控、可仿真和可灰度，在需求阶段做分级变更定义，在研发阶段做规范、测试和发布，在预发阶段做仿真回放和AB灰度；
* 事中要做到可监控、可演练、可应急，数据全链路和应急监控等各个环节都能做演练和巡检；
* 事后要做到可度量、可审计和可持续，包括事件管理、问题故障审计报告、案例学习和晋级可晋级考试等，蚂蚁每年会有一次公司级别的数据红蓝攻防，也有一年两到三次的必须参加的安全生产晋级考试的运营活动。

**6. 数据质量治理案例**（1）数据变更免疫数据变更免疫的核心目标是希望让错误代码不发布到生产。为了实现这一目标构建了几道防线：事前构建变更准入防线，将变更必须满足的“三板斧”要求、发布窗口要求等风险底线要求植入到变更准入的防线；事中构建变更灰度防线，在变更生效之前，用真实的流程去预验验证，提前发现问题；事后重点是变更监控，变更生效之后，能够持续监控变更的业务变化，有问题快速进行恢复。下面这张图，是面向发布环节研发的发布管控产品。



所有的变更在通过该产品发布都需要进行校验，类似于现在业界比较火 DataOps，将测试、灰度、仿真、监控全部纳入到流程中，做到在发布的时候自动化地进行质量监控和巡检。（2）红蓝攻防红蓝攻防的核心思路是通过故障的注入，对生产链路进行模拟攻击，发现防控体系的薄弱点。



模拟在线环境，用任务攻击和数据攻击两种方法进行攻击。在进行数据红蓝攻防的过程中需要解决三个核心问题：

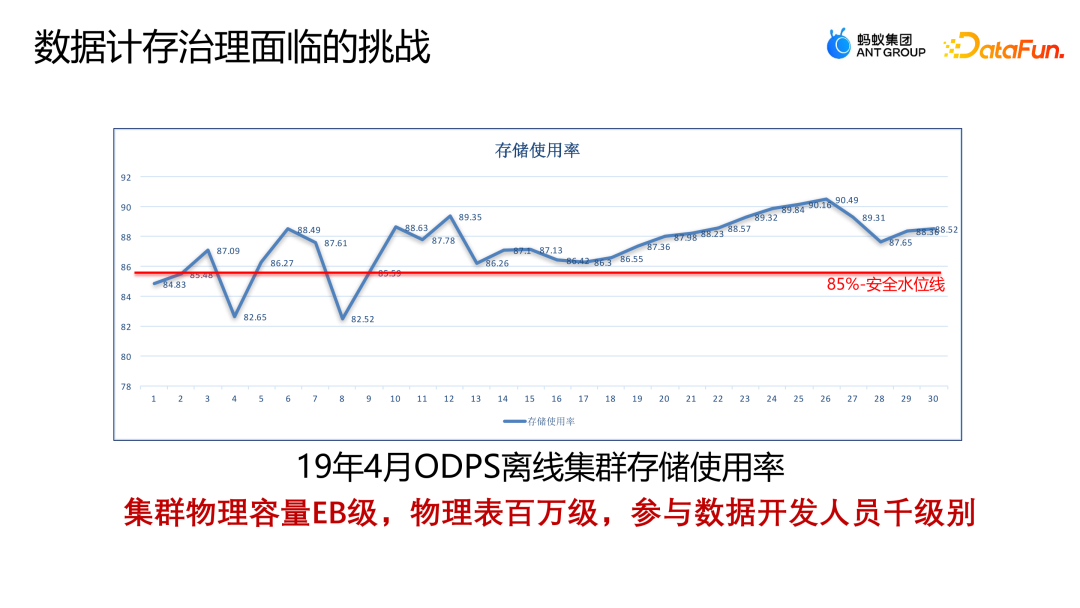
* 如何不影响生产？因为数据是一条链，上游污染了，整条数据就污染了，而且恢复成本极高。在生产环境中，构建仿真无损环境进行无差别的供给，通过攻防平台相应的数据链路在无损环境里面去植入，从而不影响生产环境。
* 如何选择攻击对象？主要选择数据入口，比如数据同步、回流任务、人群标签、有时效性保障的业务基线场景等，要重点关注有止损、有舆情的场景，比如算钱等更重要的且效果更显性化的场景。
* 如何有效地攻击？要确保所有的攻击字段能够帮助业务发现有效的生产风险，核心是通过历史故障的分析和平移，以及重大业务变更的演练。另外，在核心的攻击能力方面，构建了SQL注入等能力，以及数据大幅度波动、内容格式突然异常、资金字段错位、任务重复的回流等多种方法。

红蓝攻防在蚂蚁连续组织了四到五年，整个公司级别的红蓝攻防自动化的攻击次数达到四十多万次，推动数据质量核对规则和配置超过了五十万家，也发现了非常多的潜在问题。

**03数据计存治理**

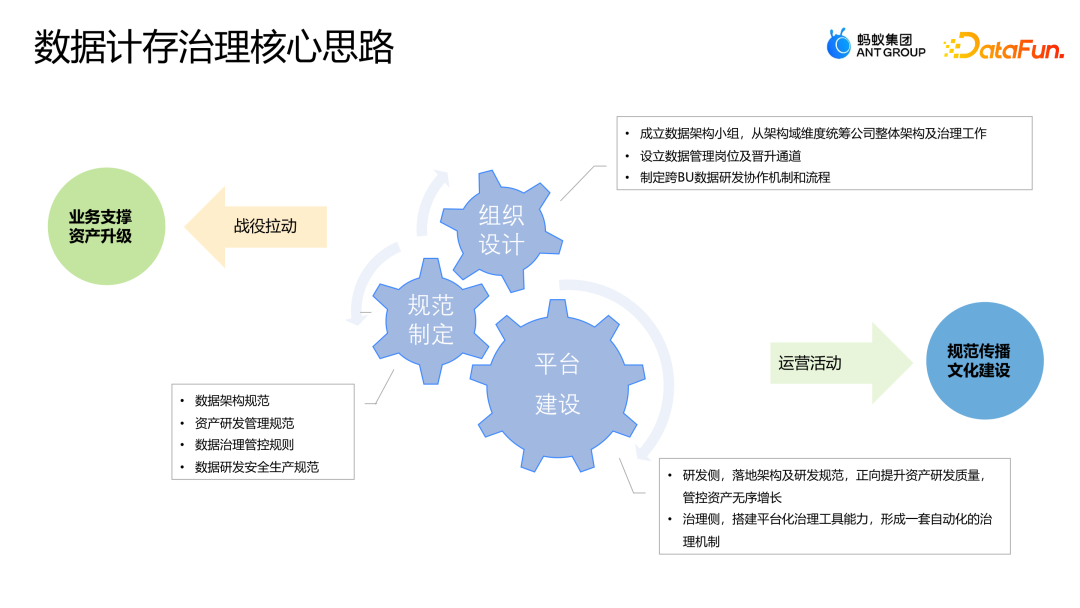
**1.数据计存治理面临的挑战**

下面这张图是2019年蚂蚁离线集群存储使用率的曲线图，安全存储的水位线是85%，一旦超过了85%就可能引发异常问题。从图中不难发现，2019 下半年集群存储使用率都在85%以上，当时出了不少安全生产问题。



计存治理会影响到安全生产。当时集群的物理容量规模已经达到了EB级，大概有几百万张数据表，参与数据研发的人员数量是几千级别的。在这样一个背景下，我们开始思考计存治理的方案。

**2.数据计存治理核心思路**

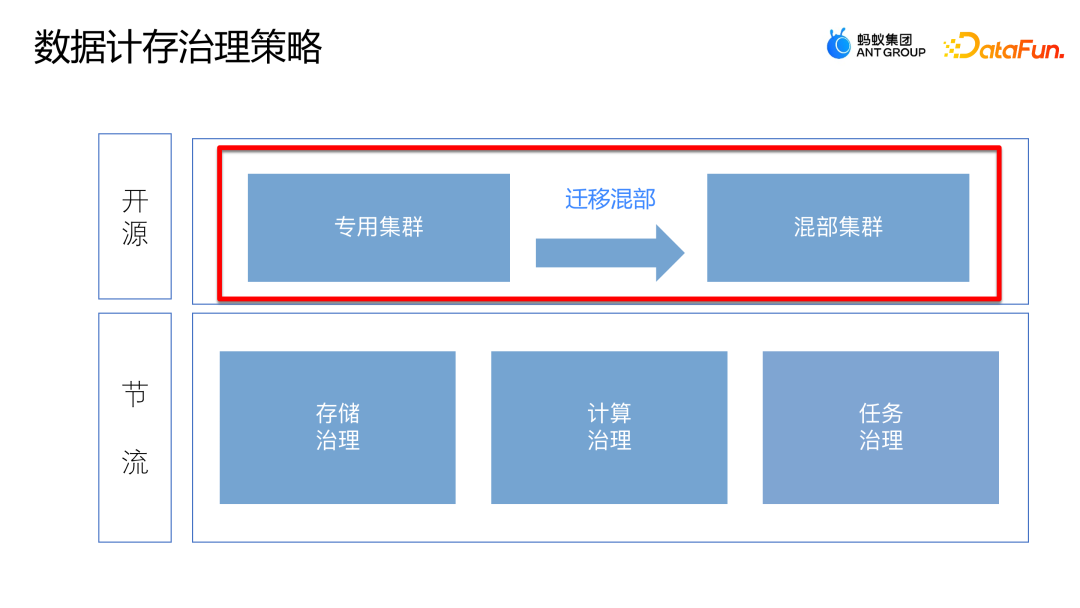


计存治理的核心思路是从组织设计、规范制定和平台建设三个方面去落地。执行的时候，通过战役拉动支撑整个业务并进行资产升级，通过运营活动进行成本规范的传播和文化的建设。

* 在组织设计层面，成立了数据架构小组。从架构域的维度统筹整个公司的数据架构和成本治理的工作。设立数据管理岗位和晋升的通道，制定研发协作机制和流程。其中，数据管理的岗位和晋升通道的设置非常关键，因为数据治理和数据管理，与数据研发，虽然都属于数据域领域，但能力与技能要求是不一样的，成长需要以不同的视角去看，所以设计了独立的晋升通道。
* 在规范制定层面，产出了蚂蚁数据架构规范、研发管理规范和数据治理管控规则。
* 在平台建设层面，研发侧正向地提升研发质量和管控资产无序增长，治理侧搭建平台化的治理工具，形成一套自动化的治理机制。

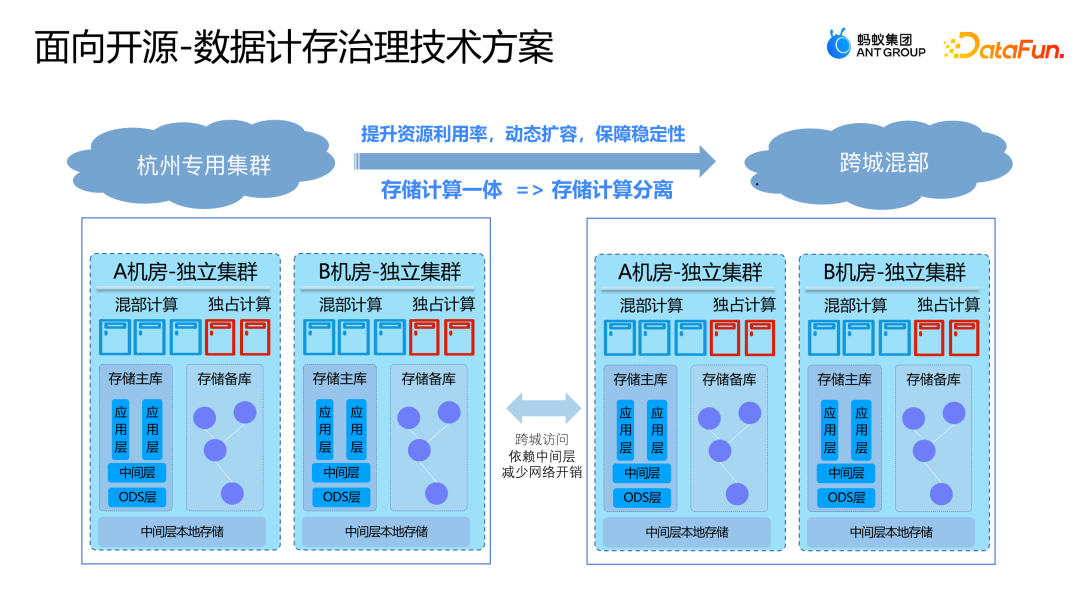
**3. 数据计存治理策略**

从开源和节流两个方面具体落地实施。



* 开源：数仓原来的资源是独享的，数仓和在线是分开的，而且数仓资源需求量非常大。在线数据库的资源使用率不高，基本在25%左右，夜间使用率可能更低，而输出储藏在夜间有非常高的计算资源需求，能不能把在线数据库空闲的资源共享给数仓离线计算呢？
* 节流：整体逻辑是数仓从任务和数据的角度尽可能去优化和节约，包含存储治理、计算治理、任务治理。

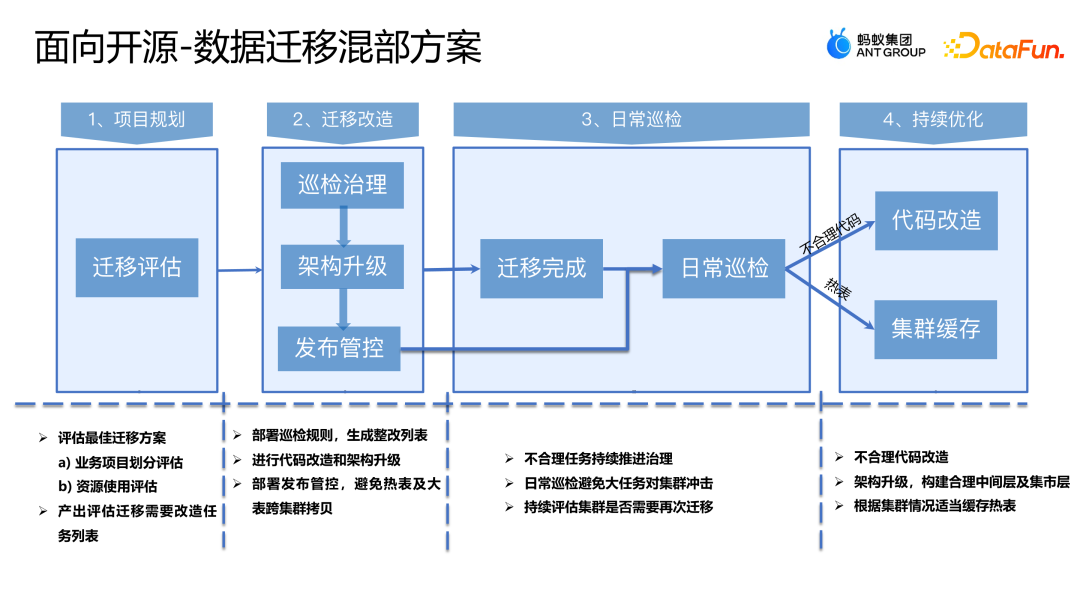
**4. 面向开源的数据计存治理方案**



以前数仓是独立的专用集群，机器、存储均独立购买。为了提供高效服务，在线应用会在本地化进行多层部署。要能与在线应用混合部署，首先要把数仓集群的架构变更到能跟在线应用混部的跨层模式，既可以提升资源利用率，又能保证稳定性。如果做成这样的“机房架构”，有两个问题必须解决：首先，如何确保数仓在高峰期不受在线资源的抢占，保证数仓高保业务在高峰期仍然可以稳定运行；其次，数仓有大量的数据交互，一旦跨层会有大量的跨层数据访问，从而带来大量的网络开销，这也会直接影响数仓的正常运行。为了解决这两个问题，核心有三件事：

* 将数仓应用层的数据访问统一收敛到数据中间层；
* 对数据中间层的热数据做跨层冗余；
* 将业务进行分级，对于高保的业务给予独占的资源，跟在线资源做适当的隔离，防止资源挤占。

存量的数据任务都是开放读取的，也存在大量的跨层访问，需要将存量也无风险迁移到整个混部的集群上来。



事前做项目规划，对业务项目划分、资源使用进行评估，产出迁移的列表；事中进行迁移的改造工作，包括部署巡检规则、进行代码改造和架构的升级、部署发布管控，避免热度及大表跨集群拷贝等；事后，做日常的巡检和持续优化，包括对跨层任务持续的监控、对不合理的代码进行改造、对热表做集群的缓存等，减少网络带宽带来的集群负载。完成混合部署后，数仓可以共享在线资源，在没有额外增加机器成本的情况下，整个数仓增加了50%的可用弹性计算资源，而且数仓任务平均等待时长降低了50%，同时，在线应用的CPU利用率也从25%提高到了40%，从全局来看，资源利用率提升非常明显。总结来说，开源的思路就是在做数据治理的时候不仅仅是只看数仓，还要将数仓的上下游及周边环节协同起来，作为一个整体来看。

**5.面向节流的数据计存治理方案**

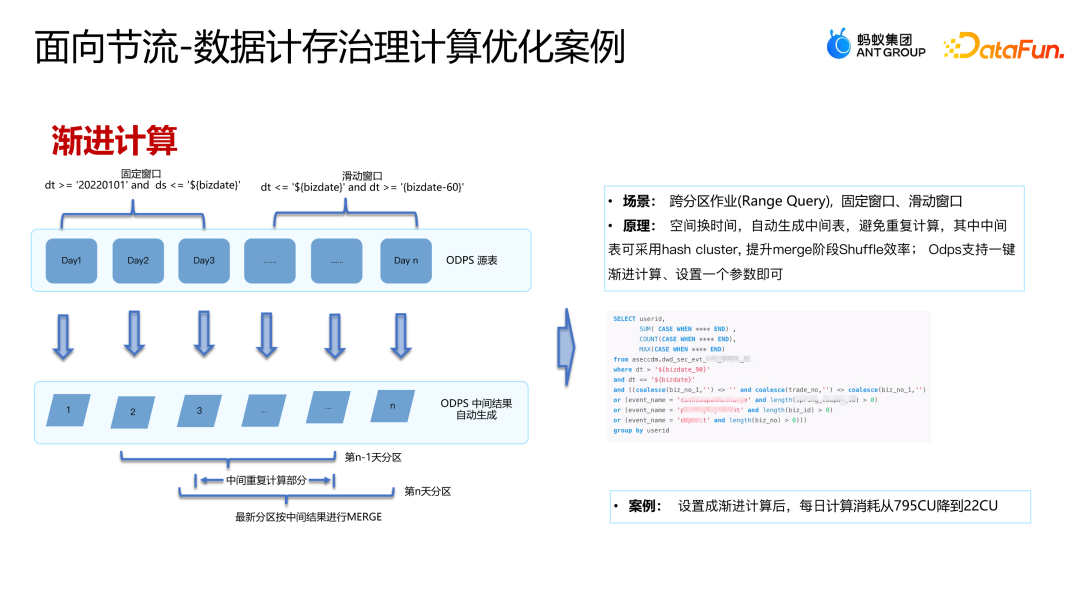


面向节流的优化可以分成几类：

* 引擎优化，比如参数优化，调度优化；
* 模型优化，比如数仓架构的链路、数仓设计、代码语法、数据压缩格式等；
* 代码优化，比如 join 的优化、UDF 的优化等；
* 资产管理优化，如果整个链路在业务上都没有应用，则考虑链路的整体下线，实现更敏捷的下线。

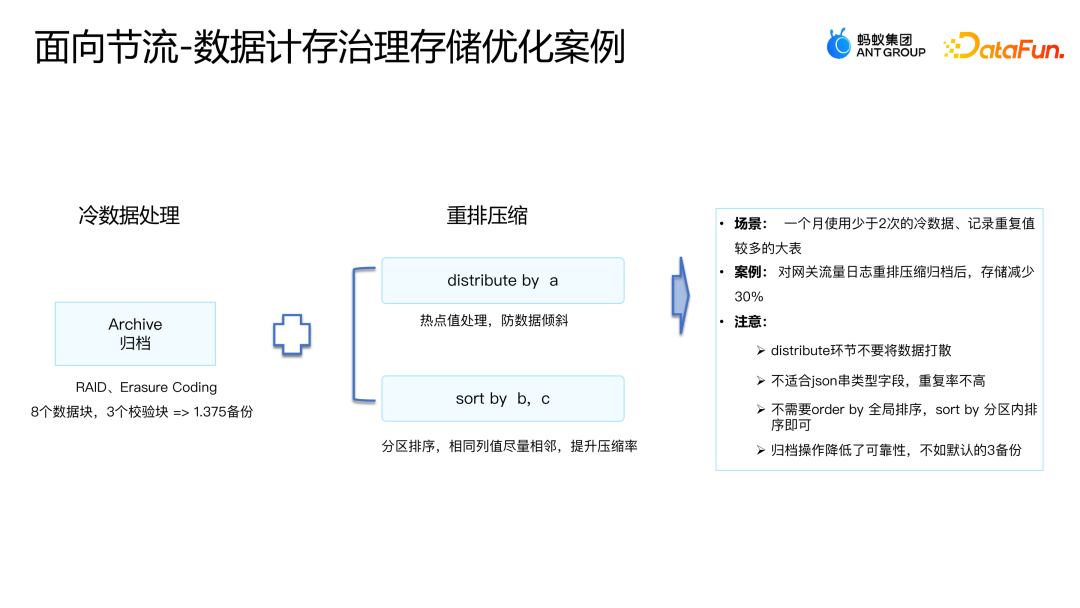
节流的整体思路就是用技术的方法提升治理自动化率，实现自动识别、归因分析、自动清理，形成常态化的管控能力。下面分享两个“小成本，大收益”的案例。

（1）渐进计算



渐进计算的适用场景是固定窗口或者滑动窗口指标计算。有固定起止日期的时间段叫固定窗口（比如年度、1月1日至今等），有固定时长的时间段叫做滑动窗口（比如近30天）。固定窗口和滑动窗口计算相同指标时有很多共性，两者在计算过程中的中间表是可以复用的，如果每次查询都重新计算就会造成计算资源的浪费。渐进计算的核心原理是“用空间换时间”，自动生成可持续滚动中间表，将中间计算的过程表保留下来，每次查询时用哈希的方式快速去读取，不用再重复计算。上图右侧是一个风控业务的案例，用渐进计算优化后，每天计算消耗从795CU降到了22CU，收益非常显著。

（2）存储归档



存储归档适用于数据查询频次不高的冷数据场景。通过对数仓数据的初步分析发现，一般访问当天数据的频率在80%左右，访问前一天数据的频率在10%-15%左右，3天前的数据很少被访问。同时，考虑到一旦对冷数据进行压缩或者重排之后，存储空间虽然会下降，但是读取时的计算性能会消耗比较大，综合考虑，将一定时间内（比如7天、30天等）未被读取的数据定为冷数据，对其进行压缩处理。当然也不是“一刀切”的方式，可以基于更精细的分析进行冷数据的定义和处理。冷数据的处理逻辑分为两类：一类叫归档，核心就是采用RAID格式的存储，用n个数据块和m个校验块的模式建设归档的能力。这样，用8个数据块和3个校验块就达到了1.375的备份，一般都是3备份。另一类是重排压缩，是distribute和sort by的结合，与电脑的磁盘整理一样，当很多空间是碎片化存储的时候，通过重排压缩把行与行之间相似的字段压缩存储。比如，相邻两行都有彭欢，存储的时候只存一个彭欢，并且告知两行都有彭欢的信息，用这种模式去优化存储。用技术的方法，不需要进行各个团队到每个人的存储或者优化，就可以带来非常大的收益。在一个案例中，网关流量日志重排压缩后，减少了约30%的存储容量。当然在进行重排压缩的时候也有一些注意事项：distribute环节不要将数据打散；不适合Json串类型字段，重复率不⾼；不需要order by全局排序，sort by分区内排序即可；归档操作降低了可靠性，不如默认的3副本。进一步，希望根据数据的冷热程度，建立自动化的识别和分级存储方案，从而实现成本的分级优化。



将数据分级成四类，在用户无感知的情况下进行自动化的数据差异化存储。

* 高频访问：热点数据，1 SSD + 3 HDD
* 热数据：访问频率正常，3HHD
* 归档数据：数据需要长期保留，访问频次低的，1.375 RAID HDD 归档模式
* 冷备存储：数据需长期保留，访问频次极低（比如监管数据等），单独建立了冷备集群，压缩比非常高，但是读取时耗费的计算资源比较高，一般是以90天的逻辑长期保留。

**04对数据治理未来的思考**

最后，分享对数据治理未来的几点思考。

* 一体化：数据在哪里治理就在哪里，随着大模型、ChatGPT、AI 的出现，以及蚂蚁自身业务的发展，目前关注在传统离线上的数据治理，未来会转变为基于湖仓一体（在线、离线、实时、图计算等）做一体化的数据治理，解决成本、合规和效率的问题。
* 价值化：数据作为生产要素，从内部的产品变为流通的商品，涉及到共享交易和开放，在数据确权价值的衡量及隐私保护方面去探索和突破价值点。
* 智能化：加入大模型做更智能的数据治理，原来是人工走向规则，接下来会探索更智能的方向。

# 内容来源：[什么是：数据格局 - 了解数据生态系统](https://zh-cn.statisticseasily.com/%E8%AF%8D%E6%B1%87%E8%A1%A8/%E4%BB%80%E4%B9%88%E6%98%AF%E6%95%B0%E6%8D%AE%E6%A0%BC%E5%B1%80-%E4%BA%86%E8%A7%A3%E6%95%B0%E6%8D%AE%E7%94%9F%E6%80%81%E7%B3%BB%E7%BB%9F/)

什么是数据格局？

数据格局指组织用来管理、分析和从数据中获取见解的数据源、类型和结构的综合生态系统。这个格局涵盖了从原始数据收集到数据存储、处理和可视化的所有内容。了解数据格局对于旨在有效利用数据进行决策和战略规划的企业至关重要。

数据格局的组成部分

一个典型的数据格局包括数据库、数据仓库、数据湖和数据集市等各种组件。每个组件在数据管理过程中都有不同的用途。数据库专为事务数据而设计，而数据仓库则聚合和存储历史数据以供分析。另一方面，数据湖允许存储非结构化数据，为数据科学家和分析师提供灵活性。

景观中的数据类型

数据格局包含多种数据类型，包括结构化、半结构化和非结构化数据。结构化数据组织严密且易于搜索，通常存在于关系数据库中。半结构化数据（例如JSON或XML文件）包含组织属性，但不适合整齐地放入表中。非结构化数据（包括文本、图像和视频）对分析提出了独特的挑战，但如果处理得当，也能提供丰富的见解。

数据来源可分为内部和外部。内部数据源包括交易系统、CRM 系统和运营数据库，而外部数据源可能包括社交媒体、第三方 API 和公共数据集。整合这些不同的数据源对于创建数据格局的整体视图至关重要，从而实现更好的分析和洞察。

数据治理与管理

有效的数据治理是数据格局。它涉及制定数据管理政策和标准，确保数据质量、安全性和合规性。组织必须实施数据治理框架来管理数据访问、使用和完整性，这对于维护对数据驱动决策过程的信任至关重要。

数据集成技术

数据集成是数据格局，使组织能够将来自不同来源的数据组合成统一的视图。通常使用ETL（提取、转换、加载）、ELT（提取、加载、转换）和数据虚拟化等技术来实现这种集成。这些方法有助于简化数据工作流程并增强数据分析和报告的可访问性。

数据分析和可视化

一旦数据被组织在分析和可视化工具开始发挥作用。这些工具使数据科学家和分析人员能够探索数据、确定趋势，并产生见解。流行的分析平台包括Tableau、Power BI和Python库，如Pandas库和Matplotlib库，它们通过可视化表示促进了将原始数据转换为可操作的见解。

数据管理面临的挑战

数据格局提出了几个挑战，包括数据竖井、数据质量问题，以及遵守GDPR等法规。组织必须通过实施健壮的数据管理实践、培养数据共享文化、以及利用机器学习和人工智能等先进技术来增强数据处理和分析能力来应对这些挑战。数据格局的未来

受技术进步和数据在商业战略中日益重要的推动，数据正在不断发展。云计算、大数据分析和物联网(IoT)等新兴趋势正在重塑组织管理和利用数据的方式。随着企业适应这些变化，数据格局将变得更加复杂，需要创新的数据管理和分析方法。

# 内容来源：思考：如何控制数据质量，怎么来提高和保证数据质量？

我们先来了解数据质量的定义，百度百科给出的解释是：数据质量（Data Management），是指在业务环境下，数据符合数据消费者的使用目的，能满足业务场景具体需求的程度。数据满足实际业务需求的程度很大程度上决定了数据质量。还记得当初J老师给我讲述的一句很核心的话：“数据就是为业务服务的，要从业务的角度来管理和处理数据”。

再来谈谈数据质量管理，数据质量管理（Data Quality Management）是指对数据从计划、获取、存储、共享、维护、应用和消亡生命周期的每个阶段里可能引发的各类数据质量问题，进行识别、度量、监控、预警等一系列管理活动，并通过改善和提高组织的管理水平使得数据质量获得进一步提高。数据质量管理是循环管理过程，其终极目标是通过可靠的数据提升数据在使用中的价值，并最终为企业赢得经济效益。

那么在数据质量管理中，如何来控制质量？数据质量控制是对数据质量进行管理和维护的过程，旨在保证数据质量，并提高数据质量。数据质量控制可以从多个维度进行，可以从以下几个方面入手：

1. 数据清洗

数据清洗是对数据进行去噪、去除错误、填充缺失值和去重等操作的过程。数据清洗可以提高数据质量，减少数据分析中的误差。

去噪是对数据中噪声信号进行滤除的过程。常见的去噪算法有移动平均（移动平均就是在数据中，按顺序依次选定一个固定长度的区间/窗口，然后计算这个窗口内数据的平均值，以平均值代替当前值。随着窗口不断向前滑动，每次都计算出新窗口内数据的平均值。）、高通滤波（高通滤波器可以滤除频率低于截止频率的信号。）、低通滤波（该算法可以保留截止频率以内的信号，而衰减截止频率之外的信号。主要用于去除高频噪声。）等。

去除错误是对数据中的错误值进行修正或删除的过程。常见的去除错误方法有数据验证（格式验证、范围验证、唯一性验证）、数据纠正（数据类型纠正、重复数据纠正、错误值纠正）等。

填充缺失值是对数据中缺失的值进行填充的过程。常见的填充缺失值方法有均值填充、中位数填充、最小值填充、最大值填充、前后值填充等。

去重是对数据中重复的记录进行去除的过程。常见的去重方法有排序去重（对数据进行排序，将相邻的相同数据进行合并。）、哈希去重（通过将数据映射成唯一的哈希值，利用哈希值来进行去重操作。把数据集拆分成小块，使用哈希函数计算小块中的数据得到对应的哈希值，将哈希值存储在一个哈希表里，检验是否已经存在该哈希值。）等。

1. 数据验证

数据验证是对数据的有效性进行验证的过程。数据验证可以通过数据校验、数据约束、数据审计等方式进行。

数据校验是对数据的有效性进行验证的过程。常见的数据校验方法有格式校验、范围校验、唯一性校验等。

数据约束是对数据的有效性进行限制的过程。常见的数据约束方法有非空约束、唯一约束、检查约束等。

数据审计是对数据的有效性进行审计的过程。常见的数据审计方法有数据审计表、数据审计报告等。

1. 数据整合

数据整合是对数据进行合并、拆分、转换、清洗等操作的过程。数据整合可以提高数据质量，提高数据分析的效率。

合并是对多个数据源进行合并的过程。常见的合并方法有内连接、外连接、全连接等。

拆分是对数据进行拆分的过程。常见的拆分方法有分区、分表等。

转换是对数据进行转换的过程。常见的转换方法有类型转换、格式转换、单位转换等。

清洗是对数据进行去噪、去除错误、填充缺失值、去重等操作的过程。

1. 数据监控

数据监控是对数据质量进行监控的过程。数据监控可以通过数据质量指标、数据质量报告等方式进行。

数据质量指标是用于评估数据质量的指标。常见的数据质量指标有准确率、召回率、F1值、精确度、噪声率等。

数据质量报告是用于记录数据质量指标的报告。常见的数据质量报告有数据质量报告表、数据质量报告图表等。

控制质量数据的方法多种多样，但不管使用什么方法，都是为了提高数据更加满足实际业务需求，真正被业务来调用，更好的服务业务。

# 内容来源：[道方图说｜解码数据合规：从基本概念、法律体系到实务探索\_企业\_处理\_主体](https://www.sohu.com/a/808175567_121123709" \l ":~:text=%E6%95%B0%E6%8D%AE%E5%90%88%E8%A7%84%E6%98%AF%E6%8C%87%E4%BC%81%E4%B8%9A,%E4%B8%BB%E4%BD%93%E6%9D%83%E7%9B%8A%E7%9A%84%E8%A1%8C%E4%B8%BA%E3%80%82)

引言：

在当前的数字经济时代，数据已成为企业的重要生产要素和核心竞争力。随着技术的进步和数据应用的普及，数据的获取、使用和共享变得前所未有的广泛。然而，伴随而来的数据泄露、隐私侵犯等风险也显著增加，迫使各国政府和行业监管机构不断完善数据保护法律法规。企业在全球范围内运营时，面临着不同法律体系下的合规挑战，不仅要确保数据的合法处理，还要有效应对复杂多变的法规要求。

本文旨在通过解析数据合规的基本概念、法律体系和实务操作，梳理企业理解和应对数据合规的核心要素。在文章中，我们试图探讨全球和中国的数据法律框架，分析实际业务场景中的合规操作，并展望未来的数据合规趋势，帮助企业建立稳健的合规体系并提升数据管理能力。

一、数据合规的相关概念

（一）数据合规的定义与内涵

数据合规是指企业在数据全生命周期（收集、存储、使用、加工、传输、提供、公开、删除等）中，遵循适用相关法律法规、行业标准和最佳实践，确保数据处理的合法性、安全性、透明度，并尊重和保护数据主体权益的行为。

数据合规与数据治理、信息安全密切相关，但有所区别。数据治理是一种企业管理框架，旨在通过有效的数据管理来优化数据价值；信息安全则侧重于数据的技术保护，确保数据在使用过程中不受未经授权的访问或攻击。数据合规则聚焦于法律和监管要求，确保企业在处理数据时符合法律的规定。因此，数据治理、信息安全与数据合规共同构成了企业数据管理的关键支柱，彼此互为补充。

数据合规与传统合规的区别。传统合规通常涉及企业在财务、税务、环境、劳动等方面遵循法律法规，而数据合规则专注于数据处理过程中的法律要求。随着数据的经济价值和隐私重要性日益增加，数据合规逐渐从传统合规中独立出来，成为企业管理中的一个关键领域。相比于传统合规，数据合规更加强调数据保护、隐私安全和跨境数据传输等特定问题。

（二）数据合规的主要内容

✦1.数据合规的法律基础

全球范围内，多个国家和地区都颁布了数据保护法律法规。欧盟的《通用数据保护条例》（GDPR）是全球数据保护领域的标杆法律，对企业处理个人数据的行为提出了严格的要求。美国的《加州消费者隐私法案》（CCPA）则侧重于保护消费者的个人隐私权，规定了消费者对其数据的控制权。中国也制定了《个人信息保护法》（PIPL）和《数据安全法》（DSL），从个人隐私保护、数据安全管理和跨境数据流动等方面规范数据处理活动。不同的法律体系有着各自的合规要求，企业必须根据所在区域和业务范围进行相应调整。

除了国家层面的法律，各个行业也有特定的合规要求。例如，金融行业需遵循《金融隐私法》，医疗行业则受《健康保险携带与责任法案》（HIPAA）的约束。这些行业法规通常要求企业对敏感数据采取额外的安全措施，并确保数据处理活动具有高度透明度。

✦2.数据合规的核心原则

（1）合法性原则。企业在处理数据时，必须基于明确的法律依据，例如数据主体的同意、合同履行或合法权益的保护。未经合法授权，任何数据处理行为都可能被视为非法，企业需对其承担相应的法律责任。

（2）正当性和必要性原则。数据收集和处理必须与业务活动直接相关，且不能超出必要的范围。企业不得为了潜在需求而无限制地收集和储存数据，这一点在GDPR和中国的《个人信息保护法》中都有明确规定。

（3）透明度原则。企业必须向数据主体清楚告知数据收集和使用的目的、范围、方式等，确保数据主体充分了解其个人信息的处理情况。在发生数据泄露或违规使用时，企业需及时通知数据主体和相关监管机构。

（4）安全性原则。企业应采取技术和组织措施确保数据的安全性，防止未经授权的访问、篡改和泄露。包括加密、数据去标识化等技术措施，都是保障数据安全的有效手段。

（5）数据主体权利保障原则。数据主体拥有访问、纠正、删除和反对数据处理的权利，企业必须为数据主体提供行使这些权利的有效机制，确保其个人隐私得到尊重和保护。

✦3.数据合规的适用范围

（1）个人信息与敏感信息

个人信息是数据合规的核心内容，指任何能够识别自然人身份的数据信息，如姓名、联系方式、身份证号等。敏感信息则包括种族、宗教信仰、健康数据、财务信息等，一旦泄露可能对个人造成严重影响，法律通常对此类数据设定更严格的处理要求。例如，《中华人民共和国个人信息保护法》规定了对个人敏感信息的特殊处理要求，企业在处理此类信息时需格外谨慎，并确保数据主体的知情同意和保护机制到位。

（2）企业商业数据

企业商业数据涉及企业的核心竞争力，涵盖商业秘密、客户数据、研发数据等。在合规管理中，虽然这些数据不直接涉及个人隐私，但其泄露可能对企业的市场竞争力构成严重威胁。因此，法律要求企业对其商业数据进行适当的保护。相关法律如《反不正当竞争法》和《商业秘密保护法》对企业商业数据的保护提供了法律基础。企业需采取技术手段和管理措施来防止未经授权的访问和使用，尤其是在数据共享和交易过程中，确保合规。

（3）大数据分析结果

大数据分析通过对大量数据的处理，可以推导出有价值的信息。然而，即使经过匿名化处理，某些情况下仍可能通过分析结果推导出个人身份，这给隐私保护带来挑战。因此，法律对大数据的使用提出了特殊的合规要求。例如，《通用数据保护条例》（GDPR）和《中华人民共和国个人信息保护法》都对大数据的合法使用作出了具体规定，要求企业在使用大数据时确保数据主体的权利得到充分保障，并采取必要的技术措施防止隐私泄露。

（4）公共数据

公共数据是指由政府及其他公共机构收集和产生的数据，包括社会经济统计数据、公共服务数据、交通数据、环境监测数据等。公共数据的开放和共享在推动数字经济和智慧城市发展中发挥着重要作用，但同时也面临数据安全与隐私保护的挑战。根据《中华人民共和国数据安全法》以及各地的数据开放政策，公共数据的开放应在确保数据安全的前提下进行。政府和公共机构需建立健全的数据开放管理机制，确保公共数据的合法、安全流通，并防范可能的滥用和数据泄露风险。同时，在公共数据的二次开发和使用中，企业和开发者应遵守数据使用的相关规定，确保数据合规并服务于公共利益。

✦4.数据合规的执行主体与利益相关者

（1）数据主体。数据主体是指提供个人信息的自然人。数据合规的关键在于保护数据主体的隐私权和个人信息安全。

（2）数据控制者。数据控制者负责确定数据处理的目的和方式，通常是数据合规的主要责任方。控制者需确保数据的处理符合法律法规，并保护数据主体的权益。

（3）数据处理者。数据处理者是受控制者委托进行数据处理的实体。尽管数据处理者不决定数据的处理方式和目的，但仍需遵守合规要求并确保数据安全。

（4）监管机构。各国的监管机构负责监督企业的数据处理活动，确保其合规操作。企业需要与监管机构保持沟通，并按要求报告数据安全事件和合规审计结果。

✦5.数据合规的实践与流程

（1）数据收集的合规要求。数据收集必须基于合法授权，企业需向数据主体明确告知数据用途，并获得明确同意。

（2）数据存储和保护。企业在存储数据时需采取严格的安全措施，确保数据的完整性和机密性。加密、访问控制等措施都是确保数据存储安全的重要手段。

（3）数据使用与传输。在使用和传输数据时，企业需确保数据使用符合收集时的声明，不得超出必要范围。跨境数据传输时，需符合法律规定的合规要求。

（4）数据共享与销毁的合规措施。数据共享需要确保各方在使用数据时符合合规要求，并签订相应的合规协议。当数据不再需要时，企业应采取适当的措施彻底销毁数据，防止其被恢复和滥用。

二、数据合规的法律体系

（一）中国大陆数据合规法律体系

中国大陆的数据合规体系是一个多层次、多维度的体系，涵盖了法律、行政法规、部门规章、地方性法规、司法解释、政策文件和行业标准等多个层面。这一体系不仅对数据的收集、处理、存储、传输和销毁等活动作出了明确规定，还为各类企业提供了具体的合规指导。以下是中国数据合规体系中的主要法律法规及相关制度。

✦1.法律层面

（1）《个人信息保护法》（PIPL）。《个人信息保护法》是中国最重要的个人信息保护法律，于2021年正式生效。该法明确了个人信息的定义、处理的合法性要求、数据主体权利及跨境数据传输的合规机制。PIPL规定，企业在处理个人信息时必须获得数据主体的明确同意，并需采取必要的技术和管理措施确保数据安全。此外，该法还对个人敏感信息（如健康、财务、位置数据）的处理设定了更为严格的要求。

（2）《数据安全法》（DSL）。《数据安全法》于2021年施行，强调数据安全的分级分类管理原则，要求企业根据数据的类型和重要性采取不同级别的保护措施。该法特别对涉及国家安全、公共利益和经济安全的“重要数据”提出了严格的保护要求，规定跨境传输需经过国家相关部门的安全评估。

（3）《网络安全法》（CSL）。作为中国首部综合性网络安全法，《网络安全法》于2017年生效，主要规范网络空间安全，包括数据处理的网络安全要求。法律对关键信息基础设施（CII）运营者提出了更严格的合规要求，要求其对涉及国计民生的核心数据和信息采取特别保护措施，并定期进行安全审查和风险评估。

✦2.行政法规和部门规章

（1）《个人信息出境安全评估办法》。该法规由中国国家互联网信息办公室发布，规定了企业在向境外传输个人信息时需进行安全评估的要求。企业在进行跨境数据传输时，需确保境外接收方的保护水平不低于中国法律的标准，并根据评估结果决定是否允许数据出境。

（2）《网络安全等级保护条例》。根据《网络安全法》的要求，中国实施了网络安全等级保护制度，要求企业根据其系统的安全等级采取不同的安全保护措施。此条例涵盖信息系统的等级划分、技术防护和监控机制，特别适用于数据处理系统较为复杂的行业。

（3）《信息安全技术—个人信息安全规范》。该规范由国家标准化管理委员会发布，作为对企业数据处理活动的技术指引。虽然是推荐性标准，但《个人信息安全规范》在实际操作中已成为企业合规的重要参考，特别是在数据收集、使用、传输和销毁的各个环节，提供了详细的操作指导。

✦3.地方性法规与政策文件

（1）地方数据保护条例。如上海市、广东省等地方政府根据自身经济特点和数据产业发展情况，颁布了地方性的个人信息保护和数据安全条例。这些地方法规在国家法律框架的基础上，提出了针对本地企业和机构的数据处理要求，特别是在数据交易和金融科技等领域，地方法规起到了补充和引导作用。

（2）行业政策文件。各行业主管部门，如中国人民银行、银保监会等，根据行业特点发布了具体的政策文件和数据管理指南。例如，《个人金融信息保护技术规范》是金融行业对个人金融信息保护的详细指引，要求金融机构在处理客户数据时确保信息的完整性、机密性和可用性。

✦4.司法解释和行业标准

（1）司法解释。中国最高人民法院和最高人民检察院曾多次发布涉及数据安全和个人信息保护的司法解释。例如，2020年发布的《最高人民法院关于审理涉及网络数据纠纷若干问题的规定》，为处理涉及数据的民事和行政纠纷提供了法律依据和裁判标准。司法解释不仅填补了立法上的空白，还为司法实践提供了参考。

（2）行业标准和合规要求。各行业协会或监管机构根据行业需求，制定了具体的数据合规标准。例如，工业互联网行业颁布了《工业数据分类分级指南》，对工业企业如何保护其商业数据提出了详细的要求；而在医疗行业，国家卫健委发布了《医疗健康数据管理办法》，明确了对患者数据处理的具体要求。

✦5.国际合规合作与标准接轨

中国在数据合规体系的构建中，也积极参与国际合作，并参考了欧盟《通用数据保护条例》（GDPR）、亚太经济合作组织（APEC）框架等国际标准。中国的数据合规框架正逐步与全球标准接轨，特别是在跨境数据传输、数据主权和隐私保护等领域，推动数据安全的国际化管理。

（二）国际数据合规法律状况

随着全球数据处理和传输的加速，各国纷纷制定相应的法律法规来规范数据的处理和保护。企业在全球范围内运营时，需要根据不同国家和地区的法律体系确保数据处理的合规性。以下是主要国际数据保护法律框架的概述。

✦1.欧盟的《通用数据保护条例》（GDPR）

（1）数据主体权利。GDPR赋予数据主体强大的权利，包括访问权、纠正权、删除权（“被遗忘权”）和数据可携权。这些权利为个人提供了对其数据的高度控制，并要求企业在处理个人数据时尊重这些权利。

（2）合规义务和罚则。企业必须遵守严格的数据处理规则，包括在数据收集、使用和存储时获得数据主体的明确同意。此外，GDPR要求企业任命数据保护官（DPO），负责监督合规实施。如果企业违反GDPR规定，可能面临高达全球年营业额4%的巨额罚款。

（3）跨境数据传输要求。GDPR规定，只有当数据接收国的法律提供了与欧盟相当的数据保护水平时，跨境数据传输才是合法的。企业可以通过标准合同条款（SCCs）、绑定企业规则（BCRs）等方式确保跨境数据传输的合规性。

✦2.美国的CCPA及其他法律

（1）加州消费者隐私法案（CCPA）的核心要求。CCPA是美国最严格的州级数据隐私法之一，赋予加州居民多项权利，包括知情权、删除权和拒绝出售数据的权利。企业必须向消费者透明地披露其数据处理活动，并确保消费者有权拒绝其个人信息被出售。

（2）各州法律与联邦动向。除了CCPA，美国其他州也逐渐推出类似的数据保护法律，如弗吉尼亚州、纽约州的数据隐私法。此外，联邦层面关于统一数据保护法律的讨论也日益加剧，未来可能出台全国性的隐私保护法律。

✦3.亚太地区的数据保护法

（1）日本《个人信息保护法》。日本的《个人信息保护法》要求企业在处理个人信息时获得数据主体同意，并在跨境数据传输时确保数据的合法使用。日本通过与欧盟达成“充分性决定”协议，使得两地之间的数据自由流动符合GDPR的要求。

（2）韩国《个人信息保护法》。韩国的数据保护法律严格，要求企业在处理个人信息时采取高标准的保护措施。韩国的个人信息保护委员会（PIPC）负责监督企业的合规情况，对违规行为有严厉的处罚措施。

（3）印度与其他国家的发展趋势。印度正逐步推出其《个人数据保护法案》（PDP），该法案借鉴了GDPR的框架，旨在对数据主体的权利进行全面保护。其他亚太国家也在加强数据保护立法，趋向与全球数据保护标准的接轨。

✦4.其他国家和地区的数据合规要求

（1）巴西《通用数据保护法》（LGPD）。巴西的LGPD是拉美地区最具影响力的数据保护法律，与GDPR类似，赋予个人对其数据的广泛控制权，并对企业设定了严格的数据处理要求。

（2）加拿大与澳大利亚的合规标准。加拿大的《个人信息保护与电子文档法》（PIPEDA）和澳大利亚的《隐私法案》都要求企业在处理个人信息时遵守透明度和安全性原则，并为数据主体提供纠正、访问等权利。

三、数据合规的实务探索

随着数据保护法律法规日益严格，企业在日常业务中需要有效落实数据合规要求。以下将详细探讨企业如何在实际操作中建立和实施合规体系，并应对各种合规场景中的挑战。

（一）企业数据合规的实施步骤

✦1.评估当前数据合规状况

企业首先需要全面评估现有的数据处理流程与合规状况。这包括对企业所处理的数据进行盘点和分类，明确哪些是个人信息、敏感信息以及其他需重点保护的数据。评估阶段应包括以下步骤：

◆数据盘点：识别企业内部的数据种类，涵盖从客户信息、员工信息到商业数据的各类数据资源。

◆风险评估：分析数据处理中的潜在风险，特别是与法律法规不一致的领域。企业应结合业务流程和法律要求，评估数据处理中的合规缺口。

✦2.制定数据合规策略与政策

评估后，企业需基于评估结果制定适合自身业务的数据合规策略，确保所有数据处理活动符合法律法规要求。

合规策略制定：确定企业合规目标，制定全面的合规方案，涵盖数据收集、使用、传输、存储及销毁的各个环节。

数据合规政策发布：将合规策略转化为企业内部政策，明确员工在处理数据时的责任和义务。企业需要向全体员工传达这些政策，确保每个业务部门理解并遵守合规要求。

✦3.组建合规团队

企业需要任命一名数据保护官（DPO）或组建专门的合规团队，负责监督和落实数据合规策略的实施。

数据保护官（DPO）任命：GDPR等法规要求某些企业任命数据保护官，DPO负责企业的数据保护事务，确保数据处理符合法规，并对数据泄露等问题进行及时报告。

跨部门合规团队：数据合规涉及法律、技术和业务流程，需要多部门协作，企业可通过组建合规团队加强不同部门之间的沟通与协作。

✦4.实施数据保护技术措施

在合规管理上，技术措施是确保数据安全和合规的核心部分。企业需要采取多种技术手段来保护数据。

数据加密与匿名化：企业应加密存储和传输中的数据，防止未经授权的访问。此外，使用匿名化和去标识化技术可以减少数据泄露的风险。

访问控制与权限管理：设立严格的访问权限机制，确保只有经过授权的人员可以访问敏感数据，避免数据滥用。

✦5.员工培训与意识提升

为了确保数据合规政策能够得到有效实施，企业需要对员工进行定期的合规培训。

合规培训：通过培训使员工了解数据合规的基本概念、企业政策以及法律义务，尤其是涉及处理敏感数据的员工需要掌握相关法规要求。

合规意识提升：除了技术人员，企业所有员工都应具备基本的数据合规意识，将合规要求嵌入日常工作流程中，防止因操作不当导致的合规问题。

✦6.建立数据合规审计和监控机制

企业需要定期对数据处理活动进行审计和监控，确保数据合规管理体系有效运行。

内部审计：定期进行内部合规性审计，检查是否存在合规缺口或数据处理中的问题。

第三方审计：企业也可以聘请独立的第三方机构对其数据处理活动进行合规性审查，以确保其数据管理系统符合法律标准。

✦7.数据泄露应急响应与处理机制

数据泄露事件一旦发生，企业需迅速采取应对措施，以减少损失并履行法律义务。

数据泄露应急预案：企业需制定完善的数据泄露应急预案，明确责任分工和应对流程，一旦发生泄露事件，能够迅速启动响应机制。

及时通知与报告：根据GDPR等法律，企业在数据泄露后应在72小时内向监管机构报告，并在必要时通知受影响的数据主体。

✦8.合规改进与持续优化

数据合规是一个持续改进的过程，企业需要定期对其合规策略和措施进行更新和优化。

政策更新：随着法律法规的变化或企业业务的调整，企业需定期更新合规政策，以确保其持续符合最新的法律要求。

技术升级：新技术的应用可能带来新的数据风险，企业应及时升级技术手段，确保数据合规的技术支撑持续有效。

（二）具体业务场景中的合规实践

✦1.跨境数据传输的合规实践

跨境数据传输涉及多个国家的法律要求，企业必须确保数据在跨境过程中符合不同地区的数据保护法律。

合规要求与实践：GDPR对跨境数据传输有严格规定，企业需通过SCCs（标准合同条款）、BCRs（绑定企业规则）等合规手段进行跨境数据传输。中国的《数据安全法》也对跨境数据传输提出了明确的安全评估要求，企业必须根据中国法律进行合规操作。

标准合同条款（SCCs）和跨境数据安全评估：SCCs是企业进行跨境数据传输时常用的合规工具，而中国的跨境数据安全评估则要求企业在进行跨境传输前，向相关主管部门提交安全评估报告，确保数据在跨境后能够得到足够的保护。

✦2.新技术（如人工智能、大数据）中的合规要求

新兴技术如人工智能（AI）和大数据在数据处理中具有巨大的潜力，但同时也带来了新的合规挑战。

AI和大数据在合规中的挑战：AI和大数据的使用涉及大量数据的自动化处理，可能存在侵犯隐私和数据滥用的风险。企业在使用这些技术时，必须确保数据的合法性，并采取措施减少算法偏见和隐私侵犯。

数据匿名化与合规技术：在大数据分析中，企业可以通过数据匿名化、假名化等技术来实现合规，确保数据主体的身份不会被轻易识别，降低隐私风险。

✦3.云计算与数据存储的合规

云计算服务的广泛应用对企业的数据存储和管理提出了新的合规要求。

云服务合规性审查：企业在使用云服务时，需要审查云服务提供商的数据保护措施，确保其符合GDPR、PIPL等法律的要求，并签订相关的数据处理协议。

云端数据保护的特殊要求：在云端存储数据时，企业必须确保数据的机密性和完整性，防止未经授权的访问和篡改，特别是对敏感数据的处理应采取加密等额外的安全措施。

（三）数据交易所的合规标准与管理

✦1.数据交易所的合规要求

随着数据交易市场的兴起，数据作为一种资产被频繁交易，这为企业带来了合规新问题。

数据交易合规审查机制：数据交易所作为数据交易的主要平台，必须对交易的数据进行合法性和合规性审查，确保数据的来源合法且不会侵犯数据主体的权益。

数据来源合法性与使用范围：企业在进行数据交易时，需要明确数据的合法来源，并确保数据的使用范围符合法律要求。

✦2.数据交易中的法律风险与防范

在数据交易过程中，企业面临许多潜在的法律风险，必须采取防范措施。

合同条款的设置与责任分担：在数据交易合同中，企业需要明确规定数据使用的范围、用途，并对各方的责任进行划分，以防止合规风险和法律纠纷。

数据交易后的合规管理：交易完成后，企业仍需对数据的使用进行合规监控，防止数据滥用和超范围使用。

（四）数据合规的挑战与应对

✦1.多国法律冲突与跨境数据合规

随着数据跨境传输的频繁，企业面临不同国家法律体系之间的冲突和复杂的合规要求。

如何在多国法律环境下实现合规：企业可以通过与法律顾问合作、使用标准化合规工具（如SCCs和BCRs）以及实施本地化数据存储策略来平衡多国法律要求。

✦2.数据技术的快速发展

数据技术的快速发展带来了更多的合规挑战，特别是人工智能和大数据技术的应用。

AI技术、大数据在数据交易中的合规挑战：随着数据技术的深入应用，企业需要解决算法偏见、数据滥用等新兴合规问题，同时保障数据主体的隐私。

技术与法律的适配问题：快速发展的数据技术往往超前于法律规定，企业需要提前布局，确保技术应用符合法律要求。

✦3.在隐私保护与数据利用之间寻找平衡

数据合规不仅仅是企业的法律责任，更与企业的发展紧密相关。

在隐私保护与数据利用之间寻找平衡：企业必须在保护数据主体隐私和利用数据创造价值之间找到平衡点。合规管理不仅能降低企业风险，还能通过提升客户信任度，增强企业的市场竞争力。

四、数据合规的未来发展趋势

随着技术的快速发展和全球数据保护法规的逐步完善，数据合规的趋势不断变化，企业需要及时调整其数据合规策略，以应对新兴技术和国际监管环境的变化。本节将重点探讨未来数据合规的几个关键趋势。

（一）技术创新与数据合规

✦1.人工智能与合规的关系

人工智能（AI）的快速发展带来了前所未有的数据处理能力，但也伴随着数据隐私和合规问题。AI技术依赖于大规模数据的分析和处理，企业在使用AI时面临多个合规挑战。

算法的透明性和公平性：AI算法可能会基于不完整或偏差的数据作出决策，导致“算法歧视”的问题。GDPR等法律要求企业在使用AI处理个人数据时，确保算法的透明性和可解释性。

自动化决策的合规性：GDPR规定了自动化决策的限制，企业不得在未经个人同意的情况下，通过自动化流程作出对个人有重大影响的决定。

✦2.区块链技术在数据保护中的应用

区块链以其去中心化和不可篡改的特性，逐渐在数据保护领域得到应用，特别是在数据交易和身份验证方面。然而，区块链也面临数据合规方面的挑战。

数据不可篡改与“被遗忘权”的冲突：区块链技术的核心优势在于数据记录的不可篡改性，但这一特性与GDPR中的“被遗忘权”存在冲突。如何在保障区块链透明度的同时，满足法律要求成为重要的课题。

去中心化管理的合规性：区块链的去中心化特点意味着没有一个单一的控制方，这给合规责任的划分带来了复杂性。企业需要在链上数据的共享和使用中确保合规管理。

✦3.隐私增强技术（PETs）的发展

隐私增强技术（PrivacyEnhancingTechnologies,PETs）在数据保护领域逐渐发挥关键作用，这些技术旨在通过创新的技术手段提升数据隐私保护能力。常见的PETs技术包括数据去标识化、差分隐私、同态加密等。

差分隐私技术的应用：通过添加“噪音”来防止个体数据的泄露，差分隐私技术已被广泛应用于大数据分析，尤其是在AI和数据分析的背景下。

同态加密与零知识证明：这些加密技术使得数据可以在加密状态下进行处理，从而实现数据隐私的最大化保护，尤其适用于金融和医疗等数据敏感行业。

（二）全球合规趋势

✦1.数据主权与全球合规标准的协调

随着数据成为国家战略资源，越来越多的国家开始实施“数据主权”政策，要求本国数据必须存储和处理在本国境内。这种政策的普及给跨国企业的数据合规带来了新的挑战。

数据本地化要求的增加：中国、俄罗斯和印度等国家要求关键数据必须在本国存储和处理，企业需要建立本地化的数据中心或选择合规的第三方云服务商。

全球标准的统一化趋势：随着数据跨境流动的频繁，国际上开始讨论建立全球统一的数据保护标准，例如亚太经合组织（APEC）的跨境隐私规则（CBPR）以及国际标准化组织（ISO）的数据隐私管理标准。这些国际协作为跨国企业提供了统一的合规框架，减少了因不同国家的法律差异带来的合规成本。

✦2.合规的监管加强与企业应对策略

全球范围内，数据保护的监管力度持续加强。监管机构正通过更严格的审查、罚款以及法律诉讼对企业的数据处理活动进行监管，这对企业提出了更高的合规要求。

监管机构对合规审查的加强：欧盟、美国、中国等地区的监管机构已经加大了对企业数据处理的审查，尤其在个人数据处理和跨境传输等方面，合规漏洞会导致巨额罚款。近年来，因数据泄露或违规处理而被罚款的企业案例不断增多。

企业可以通过以下策略应对日益严格的监管环境：

（1）投资数据合规基础设施：包括合规软件、自动化监控工具等，确保实时的合规性审查和监控。

（2）加强与监管机构的沟通：定期与本地监管机构联系，了解最新的合规要求，并在合规审查前主动披露合规计划，减少合规风险。

（3）定期合规审计：建立内部和外部的合规审计制度，通过审计及时发现并修复合规漏洞。

结语

数据合规已成为企业运营中的战略性课题，不仅是遵守法律的义务，更是维护企业声誉和市场竞争力的必要手段。在全球数据保护法规日趋严格的背景下，企业需要不断优化其数据管理流程，并通过技术创新和法律遵循来确保数据合规的有效实施。

未来，随着技术的进步和全球合规标准的进一步协调，数据合规将不仅限于法律层面的要求，更将成为企业数字化转型的重要推动力。企业应当将数据合规作为其核心战略的一部分，持续关注数据交易和数据处理中的合规问题，并建立完善的合规体系，以应对不断变化的法律环境和技术挑战。

对于企业而言，成功的数据合规不仅能有效规避法律风险，还能提升客户信任度，进而促进企业的长远发展。

# 内容来源：[数据模型设计 · Zion 帮助中心](https://docs.functorz.com/start/data_model_design/data_model_design.html" \l "%E8%AF%86%E5%88%AB%E5%AF%B9%E8%B1%A1)



识别对象

在开始设计数据模型之前，一般我们已经梳理出了业务需求清单和最终应用程序展现的高保真或低保真图。这一节，我们还是以个人知识付费平台为例，利用上一节中已经完成的产品需求思维导图以及高保真设计稿来进行数据模型对象的识别。

根据业务需求抽取出需要进行信息分析的对象。实际上，这里的对象也可以理解为在上一节中梳理出来的核心功能抽象出来的实体，例如“用户”、“课程合集”、“课程订单”

先将最主要的对象罗列出来，接着可以通过用户在使用应用程序时操作流程的视角来识别出其他实体或抽象的对象。这个思考方式是完善数据模型时最主要的技巧！在个人知识付费平台的用户视角中，我们首先可以找出“用户”和“课程合集”两个实体对象，然后用户是如何在应用上“购买”课程的呢？他通过在首页浏览课程合集内容，进入到课程合集中查看课程详情内容，课程详情内容包含了“章节”、“课程”、“FAQ”等，通过“立即购买”生成“订单”来实现课程合集的购买。 我们不要求在这一步就把所有可能涉及到的对象全部找出来，数据模型的设计是一个不断修改优化的过程，需要与设计稿相结合，有时候可能页面设计上的需求导致了我们数据模型的修改。

确定关键属性

在找出了对象后，我们先快速的为他们确定一些关键属性。实际上，在上一节细化核心功能时，我们基本上就已经确定了每个对象的关键属性。在设计数据模型时，我们需要思考如何准确命名这些对象以及其关键属性名称来区分他们。

用户：用户名、头像、邮箱、手机号码、

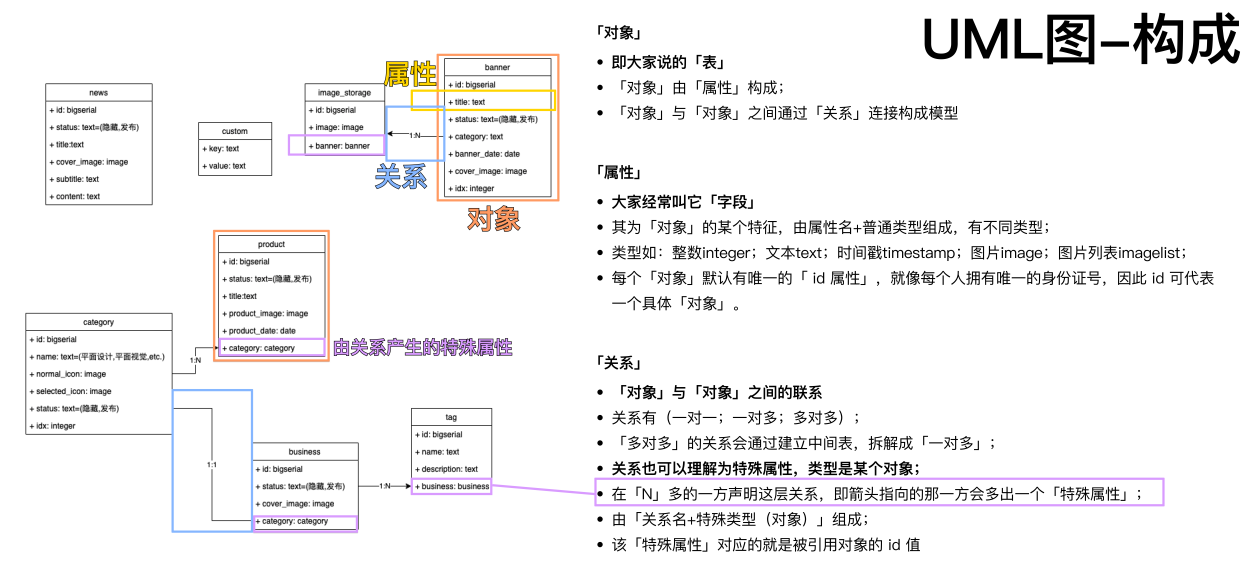
课程合集：课程合集名称、封面图、卖点内容、语种、章节名称、课程名称、FAQ

课程合集订单：购买者、课程合集、购买时间、数量、价格、折扣金额、实付金额

⚠️这一阶段罗列的属性不需要是准确的，它的类别也不重要 重要的是通过罗列这些对象的关键属性为我们更清晰的明白他们之间可能存在的关系，以及如何更合适的设计

画草图

我们可以通过绘制UML（Unified Modeling Language）图来建立一个对象关系草图来表示他们之间数据传递的方式。同时可以将上一步罗列的关键属性加入到对象当中。当然你也在纸上完成最终逻辑数据模型的，完全可以按照习惯的方式来。



绘制UML图

使用UML（Unified Modeling Language）来绘制数据模型并不是必须的

然而依托于飞书强大的协同办公能力，以及在文档中添加/编辑UML图，目前我们通常会用飞书的UML图绘制工具制作数据模型。

我们建议在绘制草图的时候就开始使用UML工具来制作，主要有两个原因：

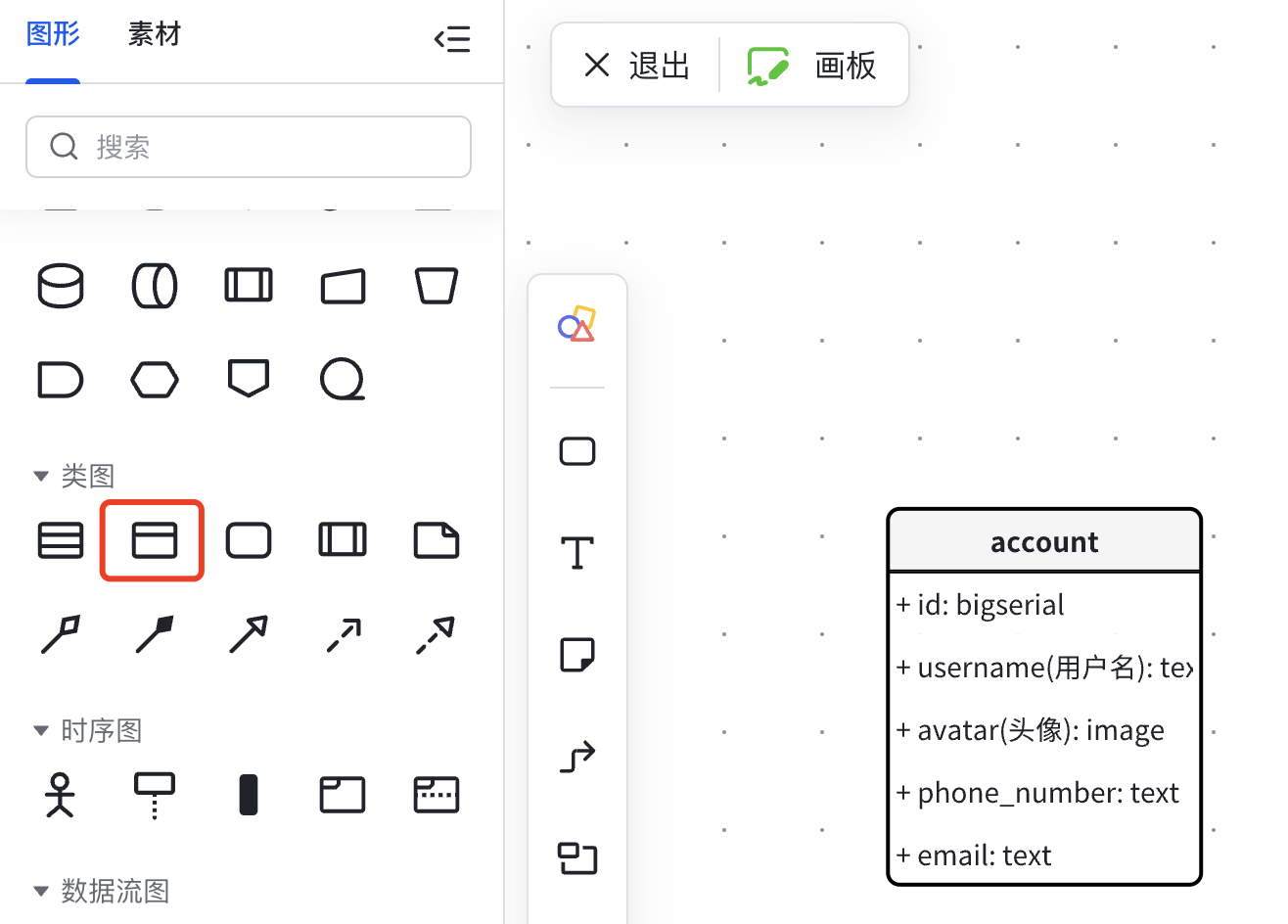
通过接下来的设计步骤最终它会形成一个可以应用在Zion上的逻辑数据模型；

我们可以将绘制的数据模型方便地同步给项目各方。

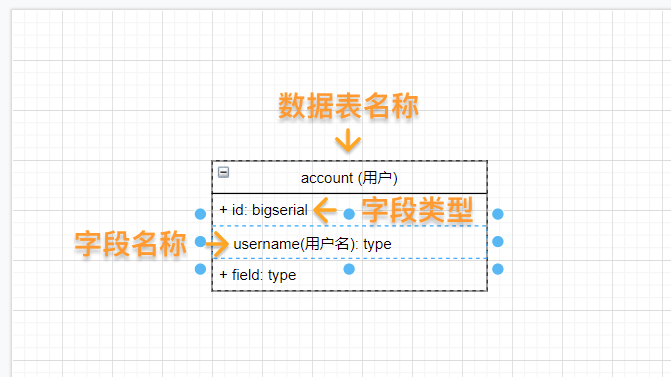
飞书文档绘制步骤

在飞书中新建一个文档或者在已有的项目相关文档中插入UML图

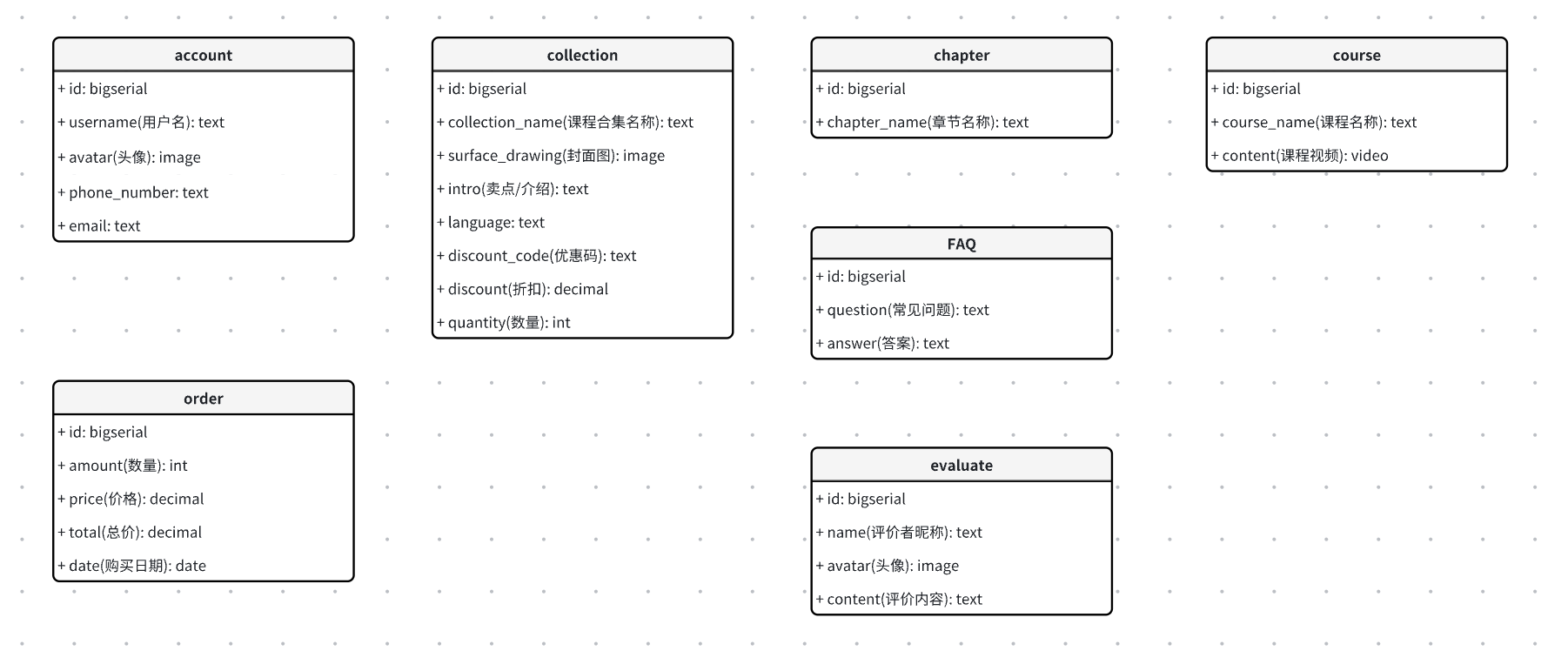
在左侧UML下找到“类图”的第二个，拖动到画布中用来添加数据模型的对象，也就是数据表



表头“account”对应的是“数据表名称”，“field”对应的是“字段名”，“type”对应的是“字段类型”



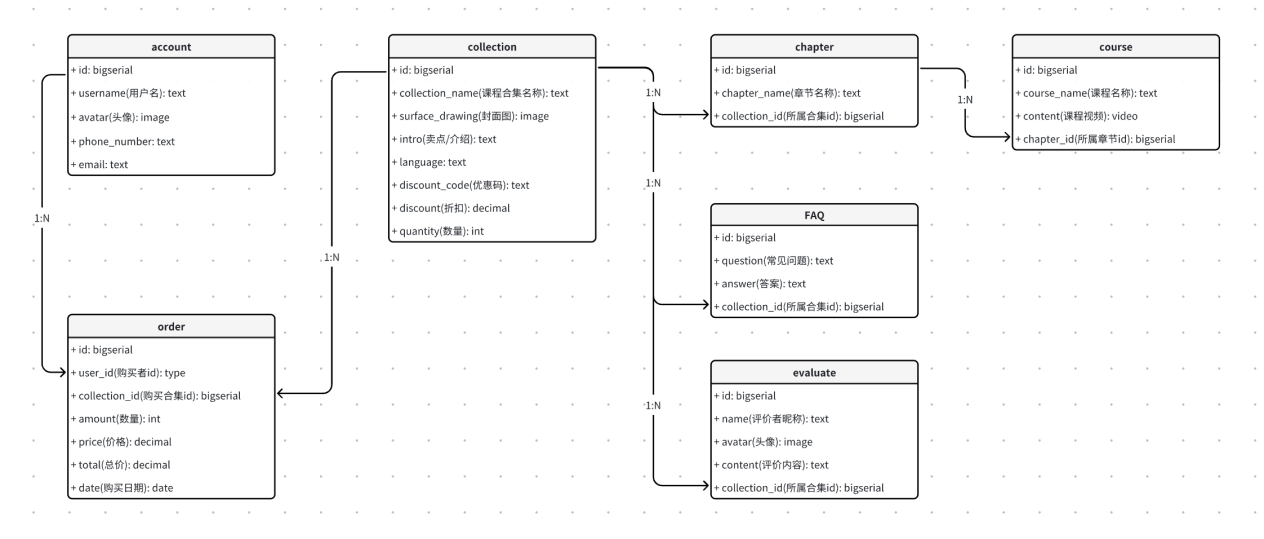
以对象“用户”举例，我们可以将数据表名称修改为“account (用户)”。 可以在名称后面加中文备注 我们通常会将数据表的第一个字段改为“id”，其类型是“bigserial”（自增长的大整数） . 将“field”改为之前找到的关键属性“username(用户名)”，type 指的是这个属性的数据类型，name一般是以文本呈现，所以是text类型。 再继续完成剩余对象数据表的添加，这里可能会有疑问，为什么合集、章节、课程、FAQ以及评价是不同的表，以及订单表中为什么没有购买者的姓名以及课程的详细信息？你可以参考[【数据模型概念】](https://docs.functorz.com/app_construction_and_maintenance/data_model_and_database/data_model_concepts/data_model_concepts.html" \t "https://docs.functorz.com/start/data_model_design/_blank)来进一步了解，也可以继续往下看



接下来，我们将为对象也就是数据表之间建立关系，建立关系的目的是让数据表之间可以相互引用数据，如果你了解了[【数据模型概念】](https://docs.functorz.com/app_construction_and_maintenance/data_model_and_database/data_model_concepts/data_model_concepts.html" \t "https://docs.functorz.com/start/data_model_design/_blank)这时候我们需要思考两个问题：

他们之间是什么类型的关系？是1：1还是1：N，还是M：N？

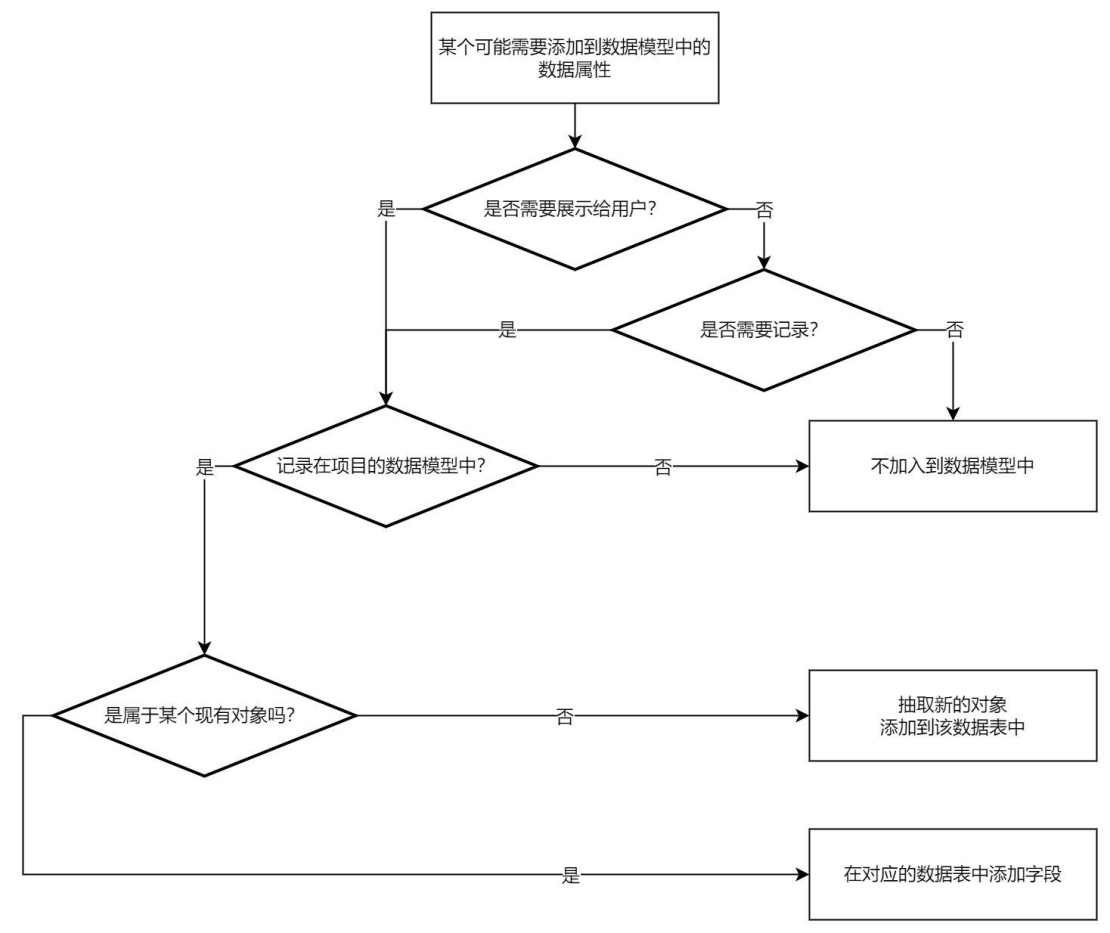
哪个对象是被引用方，哪个对象是引用方？我们来看一下“合集”和“章节”两个对象之间的关系，一个课程合集有多个章节，在章节表中，需要记录这个章节是属于哪一个课程合集，所以我们会给“合集”对“章节”建立1：N关系，“合集”就是引用方（也叫一的一方），而“章节”就是被引用方（也叫N的一方/多的一方），建立关系实际上就是在被引用的数据表中新增一个用来记录引用方的字段（这个字段在Zion中为表之间添加关系后是自动添加的，不需要自己添加），这个字段一般就是引用方的id（因为id拥有“唯一”的特性）



点击被引用对象，点击并拖动字段旁出现的蓝色箭头图标，将其拖动到引用对象的关系字段（account）上。然后可以通过点击表头位置选中数据表后来拖动调整他们的布局。我们可以通过点击关系线，然后拖动关系线上的蓝色航点来改变它的走线。双击关系线上的非航点位置，为其添加关系类型的描述，这里修改为“1:N”，然后在右侧编辑窗口打开它的背景色，这样就不会被关系线遮挡了。 至此你已经学会了在UML工具上绘制数据模型所需的所有基本技能了，并且在绘制数据模型时，你也能够根据需求来梳理好对象之间的关系。

找出数据属性

为各个对象添加所需的属性，完善数据表的字段。我们同样可以利用用户操作流程的角度结合设计步骤第二步【确定关键属性】来思考，找出现有对象外其他业务需要添加的数据。思考这个问题可以发现在步骤一中遗漏的对象。



回到个人知识付费平台中的几个例子：

用户有必要知道订单的状态，因此需要添加“订单状态”字段。

用户虽然只看到上线的商品，但是管理者需要对课程合集的状态进行管理，不再售卖的课程合集需要下架，并且可以在修改商品状态后及时的反馈给用户，因此需要添加“状态”字段# 分配属性 将上一步骤中的数据分配到对象中并且添加字段的类型以说明数据的业务含义。 这意味着我们可以通过数据表中添加的字段，反向来检查是否符合业务的需求，以及应用程序功能的满足情况。 另外，数据表的关系字段是否正确的添加在了引用方的数据表中？新属性的添加之后通常会使我们更清晰的明白数据表之间的正确关系 关系字段的命名规则： 通常情况下我们在引用方中添加的关系字段用被引用方的数据表名称表示。 而且对于1：N关系往往会采用名词复数的形式。 他是字段名称来表示是否是关系字段的提示，同时也代表了这个字段可以重复出现某一个数据。 当然也有很多时候我们会自定应关系字段名称来更好的表达他对应的含义，比如“分类”的自关联“父级分类”字段。 然后关系字段的类型填写被引用的数据表名称。 比如： “媒体库”中引用商品的关系字段更合适的字段名是“products(商品)：product” 在此基础上完善UML图的制作

校验完善

最终确定数据模型并验证其准确性。 这里的验证不仅仅是作为数据模型的制作方自己，还需要通过制作的逻辑数据模型和业务需求方确定页面设计与功能的实施是否能实现，有没有可以优化或调整的内容。 另外不要忘记数据模型是一个不断修改优化的过程，通过制作数据模型往往可以得到对业务逻辑更深入的理解。 通过以上步骤我们会得到一个理想的逻辑数据模型，接下来可以在 Zion 实现项目物理数据模型的制作。 值得一提的是，以上的设计步骤并不一定是线性的，实际上很多时候我们会交替进行，反复优化得到最终的数据模型。这些过程在放到具体的数据库工具之前，也就是物理数据模型制作之前都是合理且必要的。 而一旦在 Zion 上制作完数据模型并且将数据进行绑定和上线之后，修改数据模型将会带来很多不必要的麻烦。

# 内容来源：Akso第一个工作记录本

1. 同城双活：同一个城市有两个数据中心
2. 异地容灾：

# 内容来源：<https://mp.weixin.qq.com/s/eZdUm0N8xrKtML2bSLmxXA>

随着公司从初创企业成长为成熟企业，对数据团队的要求也成倍增加。能否有效地收集、处理、分析数据并从中获取见解，决定了一家公司是能够继续扩张，还是只能苦苦挣扎跟上竞争对手的步伐。

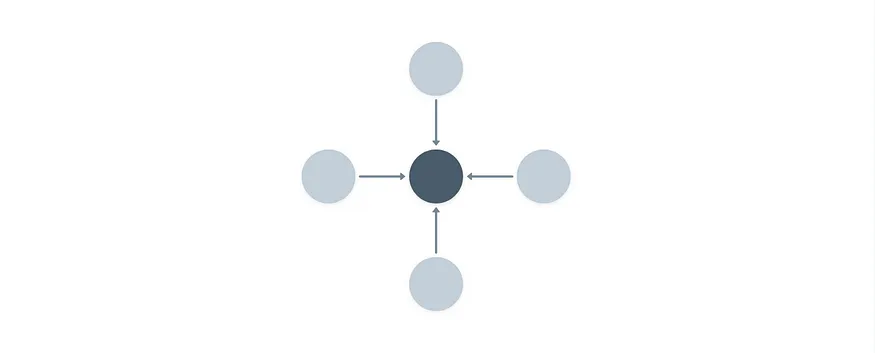
然而，组建一支能够随公司发展而扩展的数据团队需要深思熟虑和规划。这需要了解业务需求、对团队结构和角色分配采取战略性方法，并致力于在此过程中培养数据驱动的文化。

在本文中，我将就如何适应公司不断变化的需求提供一些关于团队结构和组成的想法。从早期初创公司的精益、集中式团队到大型企业的复杂、分布式结构，我将研究需要哪些角色和职责来帮助您的数据团队跟上业务增长的步伐。

无论您是希望组建首个数据团队的创始人，还是希望优化现有结构的数据领导者，本文都将提供实用建议，帮助您成功开展业务。

一 初创企业模式：集中式结构

对于初创公司来说，集中式数据团队结构通常是最高效、最具成本效益的方法。在这种模式下，所有数据人员都属于一个为整个公司服务的团队，负责数据收集、处理、分析和报告。这种结构通过将专业知识集中在最需要的地方来最大限度地利用有限的资源，并确保整个组织的数据处理实践保持一致。通常，可以从提供全方位支持的商业智能分析师开始。



优点：

统一治理：集中式团队确保数据治理和质量控制在整个组织内统一应用，减少不一致和错误。

资源管理效率：由于技能和工具集中汇集，单一团队可以更好地管理有限的资源。

简化沟通：所有与数据相关的任务由一个团队处理，沟通得到简化，从而可以简化项目管理和决策过程。

缺点：

可扩展性问题：随着组织的发展，集中式团队可能难以满足各个部门日益增长的需求，从而可能会减慢响应时间。

有限的业务部门重点：集中式模型可能导致数据团队与特定业务部门需求之间的脱节，因为中央团队可能无法完全理解或优先考虑各个部门的情况。

瓶颈的可能性：集中式团队可能会被大量的请求压垮，从而导致数据处理和分析的延迟和潜在的瓶颈。

角色详细信息

商业智能分析师

商业智能分析师专注于将原始数据转换为有意义且有用的信息以进行业务分析的工具和技术。他们开发和管理 BI 解决方案，提供报告并进行复杂的数据审查，以提高业务效率和生产力。

早期发展阶段：扩大集中式结构

当您的业务进入早期增长阶段时，在允许扩展的同时保持集中式结构非常重要。确保跨团队学科的协作并鼓励知识建设。可以引入数据工程师、数据分析师和数据科学家（如有必要）等新角色来支持组织不断增长的数据需求。但是，随着公司的发展，如果数据需求变得过于多样化和庞大，集中式模型可能难以有效扩展。

角色详细信息

数据工程师

数据工程师负责设计、构建和维护用于数据生成、处理和分析的架构。他们开发和构建数据管道以连接不同的数据源，确保分析师和科学家可以访问和使用数据。他们还管理和优化数据存储解决方案。

数据分析师

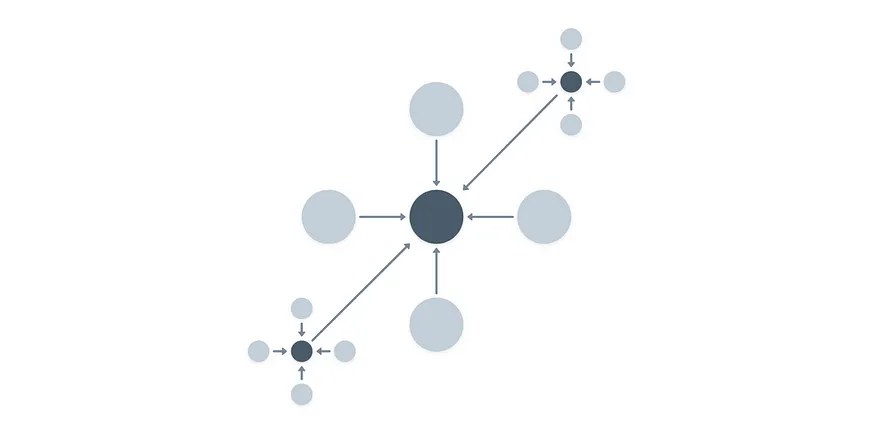
数据分析师解读数据，为业务决策提供切实可行的见解。他们使用统计工具分析数据集、创建报告并开发仪表板，以易于理解的形式可视化复杂数据。他们的工作对于帮助组织了解绩效指标和识别趋势至关重要。

数据科学家

数据科学家应用先进的统计技术和机器学习算法来模拟复杂的行为、趋势和推论。他们支持预测分析和决策过程，帮助根据大量数据集预测结果。他们经常与业务利益相关者密切合作，将数据洞察转化为战略行动。

二 中型企业模式：向混合结构过渡

当您的业务达到中型阶段时，过渡到混合结构可以在中央控制和部门灵活性之间实现必要的平衡。在这种模式下，集中式数据团队负责核心数据基础设施和治理，而其他数据角色则分布在不同的业务部门。这种结构可以更快、更专业地响应部门特定需求，同时保持标准和治理的集中性。可以引入机器学习工程师、数据治理分析师和技术业务分析师等新角色。清晰的沟通渠道和明确的角色和职责对于避免冲突和重复工作至关重要。



优点：

平衡方法：将集中治理的效率与分散团队的响应能力相结合。核心数据基础设施和合规性由集中管理，而分析和运营数据任务则由业务部门内的嵌入式团队处理。

灵活性和相关性：通过允许分散的团队专注于特定的业务领域，混合结构可确保数据解决方案与部门需求和挑战紧密结合。

可扩展性和适应性：这种结构可以随着组织的发展更有效地扩展，适应不同业务部门的需求，而不会给中央团队带来过重的负担。

缺点：

复杂的协调：需要强大的协调机制来确保一致性并防止数据孤岛，这可能会使不同团队之间的数据管理复杂化。

资源重复：跨团队可能会出现工具和技能重复，从而可能导致效率低下和成本增加。

治理挑战：在分散的团队中维护一致的数据治理政策可能具有挑战性，需要明确的政策和定期的监督。

角色详细信息

机器学习工程师

机器学习工程师根据业务需求设计和实施机器学习应用程序。他们的职责是创建算法，使计算机无需明确指令即可执行特定任务。他们与数据科学家密切合作，将算法集成到可扩展、可投入生产的系统中。

数据治理分析师

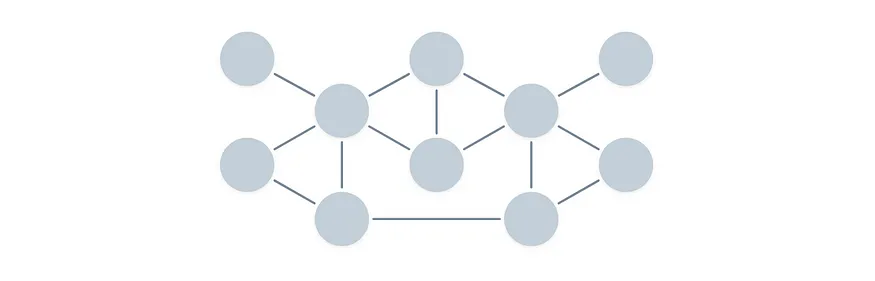
数据治理分析师负责制定和实施数据标准和实践，以管理和确保整个组织的数据质量。他们监督数据隐私、合规性和安全政策，努力确保数据的准确性、可用性和安全性。

技术业务分析师

技术业务分析师(Tech BA)在弥合 IT 解决方案与业务问题之间的差距方面发挥着至关重要的作用。他们帮助设计系统并根据业务目标提高 IT 流程的效率。

三 大型企业模式：采用数据网格方法

对于大型企业，数据网格方法可以支持复杂且大规模的数据操作。在这种模型中，数据被视为产品，跨职能团队负责不同的数据产品。这些团队是自主的，但遵循由中央管理机构设定的通用互操作性框架。这种方法通过授权团队开发适合其特定情况的解决方案来鼓励创新。可以细化角色以产生战略影响，包括添加数据产品所有者和数据架构师。一个拥有强大治理模式和跨团队高水平数据素养的成熟组织对于数据网格方法的成功至关重要。



优点：

高度可扩展：数据网格专为大型组织设计，可适应复杂且大量的数据操作，支持众多自主团队。

增强创新：通过将数据视为产品，鼓励团队独立创新和优化其数据产品，从而可能获得更快、更有针对性的数据解决方案。

对业务需求的响应能力：自主团队可以快速响应其特定领域内的变化和需求，从而提高业务敏捷性。

缺点：

管理的复杂性：管理数据网格涉及相当大的复杂性，特别是在确保不同数据产品之间的互操作性和一致性方面。

高级数据治理：需要高级级别的数据治理，以确保整个组织尽管具有自主权，但数据质量和合规性不会受到损害。

对专业知识的需求更高：每个团队不仅需要具备数据管理方面的高水平专业知识，还需要具备理解业务环境的专业知识，而维持这些专业知识很有挑战性。

角色详细信息

数据产品负责人

数据产品负责人(DPO)是组织内的关键角色，尤其是在数据驱动型产品和服务是业务战略核心的情况下。DPO负责管理数据资产和项目的生命周期，从构思到开发、发布和持续改进。他们是数据团队和业务部门之间的重要纽带，确保数据产品能够创造价值并与总体业务目标保持一致。

数据架构师

数据架构师负责设计、创建、部署和管理组织的数据架构。他们定义不同数据实体和IT系统如何存储、使用、集成和管理数据。他们还制定数据管理标准，包括确保数据库和数据存储技术符合公司需求和监管要求。

扩展数据团队的其他注意事项

无论处于哪个增长阶段，有几个因素对于成功扩大数据团队都很重要：

强大的数据领导力和明确的治理政策对于将数据作为资产或产品进行管理至关重要。

投资正确的技术堆栈和支持所选组织结构的工具可以提高效率。

不断发展和培训所有团队成员的数据素养和数据相关技能对于维持高效的数据团队至关重要。

四 小结

为数据团队选择正确的结构是一项关键决策，它应与业务当前需求和未来增长计划保持一致。随着组织的发展，调整和发展数据团队结构以有效支持不断变化的数据格局至关重要。通过了解每种结构的优势和注意事项并实施必要的角色和实践，可以成功扩展数据团队以推动业务增长和决策。虽然数据团队没有“一刀切”的方法，但我希望这可以帮助规划公司发展的下一阶段的数据组织。