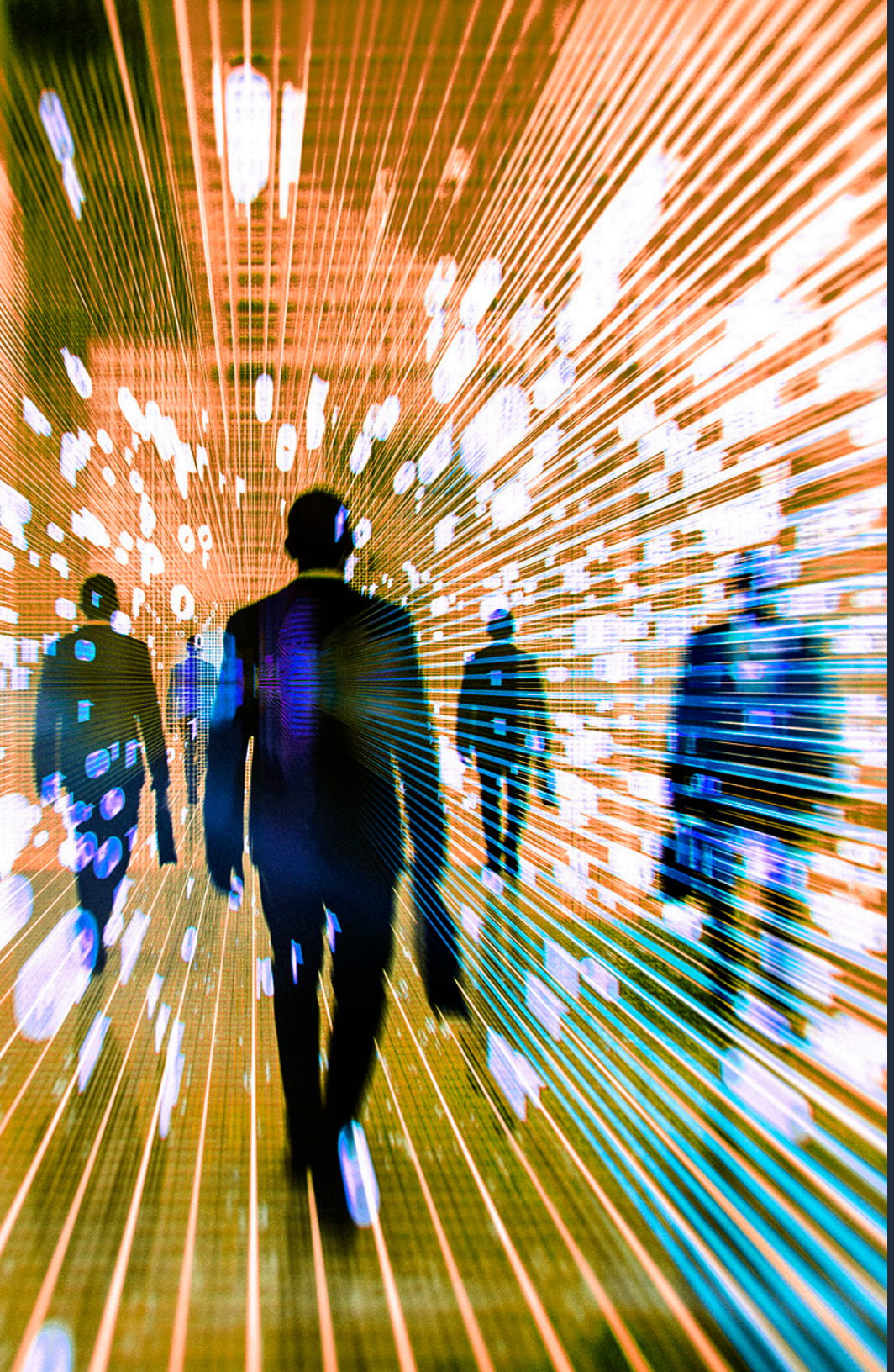
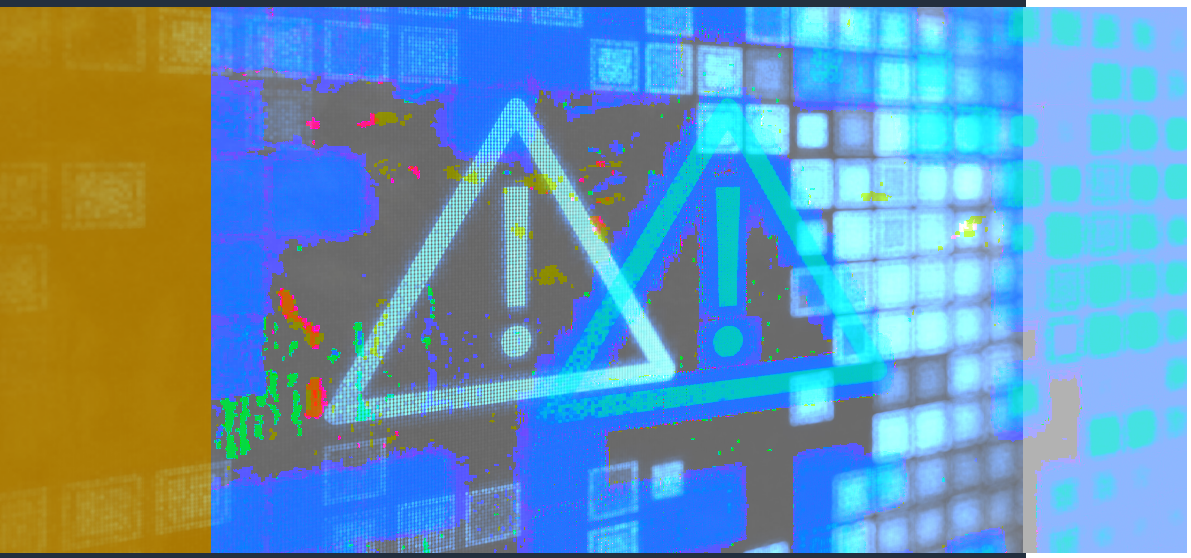


成为数据驱动型组织

Joe Chung，Amazon Web Services
的企业策略家与宣传官

每个公司都存在数据问题



想象一下……

Excel 周报表已出炉，并发送到您的邮箱中。在浏览报表时，您发现财务数据中存在自己无法理解的异常，尽管报表中已提供了数据透视表，让您至少可以在一定程度上了解到更为详细的数据。您询问您的运营分析师，这是怎么回事。对此，您的分析师回答说，“我也不清楚。我需要查看一下。”

第二天，您的分析师告诉您，出现异常是因为制造厂商的生产力下降所致。

> “这有什么影响，”您说。“你问一下 HR 病假是否会影响生产力数据？ 或者厂商的时间捕获应用程序是否有问题？”
>
> 您的分析师回答，“获取该数据并将其合并到财务数据需要一周的时间。”
>
> “你能不能给我发一份从 ERP 和时间应用程序转存的数据，我自己来处理？”
>
> 分析师回答道，“我没有数据访问权限，要想获得权限，我需要提交正确的认证票据，而这需要数天时间。”

如果您对这种对话情景并不陌生，则说明您的组织遭遇了大数据问题。

您的第一反应可能是，这是商业智能流程和工具难题，它从一开始就一直困扰着各大组织，而不是真正的大数据问题。商业分析、商业报告和商业智能之间并不需要进行一场宗教大辩论，事实上**每个组织都正面临着大数据问题。**随着人工智能和 Machine Learning 的成熟发展，企业现在比以往任何时候都更需要较好地处理自身拥有的数据，并利用这些数据让自身发展成为**数据驱动型组织。**

让企业成为数据驱动型组织的心智模型类似于人体的神经系统。神经末梢延伸至我们的身体中，向脊髓传递感知信号，然后由大脑处理和执行。心智模型是一个由数据架构模拟的模型，它支持在企业内外的任何地方实时接收、处理和存储数据。信号经过实时处理，然后通过 Machine Learning 算法执行。遗憾的是，有太多组织认为有了这些新功能更好，但只适用于专门的数据场景，还有些组织则试图将传统的商业智能平台重新定义为“数据湖”。

数据功能障碍

我们大多数人都认为大数据问题是一个关于数据体量的问题。事实上，除数据体量外，每个组织还面临着其他数据问题，这些问题被以多种不同的方式掩盖。下面是我观察到的一些常见的数据功能障碍实例：

孤立数据和被丢弃的数据

首先，许多组织并未意识到很多有趣的数据被丢掉或无法访问。例如，应用程序中的用户活动数据（以及用户如何关联使用其他应用程序的数据）、托管应用程序的基础设施的遥测数据和与当前表架构不兼容的旧版架构中的数据。

其次，数据分散在多个应用程序和数据仓库中。一个应用程序可能不“大”，但，加在一起就可能变大。因此，当企业需要对多个来源的数据进行分析时，它将变得非常具有挑战性。这是因为分散数据会带来访问权限问题。每个存储数据的地方都有自己的访问规则和规范，这将使数据访问具有挑战性。

“

每个组织都面临大数据问题，它们只是以不同的方式被掩盖了。”

低保真数据

传统的企业系统大多只处理和捕捉最终状态，通常只报告一小段时间内的数据快照。此外，只能批处理，而非实时处理数据。数据在批处理窗口之间可能会发生较大变化，而传统系统通常设计成将这些变化数据作为临时数据丢弃，因为它们的处理速度无法跟上数据的变化。

开平方表中的四舍五入数据

许多企业意识到，有些数据并不完全适合采用传统数据库存储技术进行存储（例如，图像、传感器数据等）。此外，分析数据和获取洞察力的方式也有很多种。例如，在推出新的分析计划时，您可能会发现没有哪一种单一的报告或可视化解决方案可以满足所有用户的需求。您可能需要考虑通过 API 获取经过算法处理的洞察力，而在应用程序中，则可通过使用 JavaScript 框架（如，D3.js）定制可视化小部件，以及通过 Tableau 商业智能门户及其他可视化解决方案获取洞察力。

散乱数据

企业系统中可不欢迎散乱数据，这也是为什么在存储数据前需要设计表单、规则及其他验证方式来确保数据的纯净。但是，有些最有趣的数据可并不那么纯净。在访问非结构化或对象数据时，您可能会发现噪声数据。正如电子噪声一样，你也可以通过过滤、增强和放大机制来获取想要的数据。而这又让许多企业担心将数据发送至专用日志聚合，会导致其安全性和监控工具成本增加。而在大多数情况下，通过过滤掉大部分无用的日志数据，可降低成本。

如果您在上述任何一点上遇到困难或产生共鸣，那么您的组织就应该重新审视所用的分析方法和架构了。每个企业都有机会部署适合自身用途的分析解决方案（存储、处理、查询、分析、演示等），以应对当前在业务和 IT 中遇到的挑战。

现代分析平台可以实现关键业务洞察

如果您准备利用现代分析平台处理大数据问题，那么您的合理预期是什么？ 从技术角度来看，以下就是其可能实现的目标和方式。

> “做数据驱动型决策时，需要访问许多不同类型的信息。”

访问我想要的任何数据

做数据驱动型决策时，需要访问许多不同类型的信息。飞行员依靠飞机上的仪表来获取重要的飞行信息，如高度、空气速度和燃料消耗量。但想象一下，如果这些仪表并未放在一个地方。这样，飞行员也许需要步行到后舱或通过无线电获取这些信息，或者更糟糕的是，他们在访问数据时还须先获得许可权限。不幸的是，这就是当今企业的真实写照。

前瞻性组织已将这一标准抛之脑后，它们将系统中现有的数据转存到一个地方（即，数据湖）。虽然有些公司存储大量同类型数据，但是，越来越多的公司开始创建含有多种类型不同来源数据的企业数据湖。

亚马逊、雅虎和 Facebook 等互联网规模公司，在 21 世纪初就开始意识到，关系型数据库技术在可扩展性和性能方面已经达到了极限。为此，亚马逊推出了一项名为 Dynamo 的技术，Dynamo 是一项高度可用和可扩展的 key-value 模式存储技术，如，NoSQL /非关系型数据存储技术。之后，亚马逊不断完善并利用 Dynamo 技术创建了这类服务，如 Amazon S3 和 Amazon DynamoDB 等。Amazon S3 对希望创建数据湖的企业很有吸引力，因为它能够存储许多不同类型的数据，并且成本低廉。当然，还有其他技术解决方案，包括 Hadoop，但所有数据湖解决方案的一个重要特征都是它们能够存储任何类型的 PB 级数据，且成本低廉。

变化响应性

业务系统和数据一直在变化，但报告或分享这些信息的系统常常最后才变。有多少次您被告知，修复数据仓库和报表中的数据需要 6 个月或更长时间？ 或者来自源系统的变化数据还未传输到报表系统中，并且由于批处理的关系，还需要几天才能完成变更？ **数据的访问速度决定了决策速度**。因此，我们应该期盼现代分析系统能够以接近实时的速度处理和报告数据，并响应上游数据源的变化。

> “数据的访问速度决定了决策速度。”

第一个关键驱动因素是如何通过 Amazon S3 或 Hadoop 等大数据技术存储数据。更改关系型数据库的一大阻碍就是修改存储架构或定义数据存储方式。在架构修改完毕前，数据无法存入数据库或传输中断。与“您需要适应我的结构”方法相比，采用 Amazon S3 等基于文件或对象的技术则无需担心数据结构，数据可以保持原样。

另一个难题是，在给定时段只有一个构架处于活动状态。我相信大家已经看过名为“2015”和“2016”的数据库表单，它们并不够理想。大数据技术是一个基于读取的架构，这就意味着数据结构在获取数据时就已使用，而不是根据数据存储方式推导后使用。这就意味着，源系统数据变化对企业而言并不是一件大不了的事情。

第二个推动因素是流媒体技术，如 Amazon Kinesis 和 Apache Spark。大多数企业通常每天一次，大批量转存数据。流媒体技术支持以较小数据块的形式大规模存储数据。例如，扬声器制造商 Sonos 每周利用 Amazon Kinesis 处理 10 亿个事件。您再也不用等到每天批处理完成后，才能了解自己的业务状况了。

我想要交互洞察的地方及方式

目前，企业要想了解近在眼前的信息，还需要越过重重关卡。它们可能需要翻遍收件箱，找到附件中的报表。或者，只能登录报表系统下载 PDF，然后将数据复制粘贴到 Excel 中才能看到。我们要避免用户在获取他们所需的数据和洞察力时遭受这样可怕的经历。用户的心声应该是：在正确的时间用正确的工具以正确的形式提供数据。

Tableau、Amazon QuickSight 及其他软件在这方面做得更好，它们通过数据交互提升了用户体验。然而，我发现，大多数企业需要同时使用多种工具以满足用户需求。它们可能会将 Amazon QuickSight 嵌入商业智能门户，通过电子邮件发送 Tableau 工作簿。AWS 通过边建边用 (Pay-as-you-go) 模式提供了多种数据存储和商业智能工具。这使得各组织能够尝试使用多种不同的商业智能工具，而无需在基础设施和许可上投入过大。

我们要避免用户在获取他们所需的数据和洞察力时遭受这样可怕的经历。”

数据科学家是您组织中不容忽略的人物。Jupyter Notebooks 在数据科学圈非常流行，它涉及内容管理、代码执行和可视化部分。它是一款功能非常强大的工具，用于共享知识、记录和执行 Machine Learning 算法。Amazon SageMaker 笔记本是一个托管平台，它可以帮助您及您的数据科学家分担繁重的任务。

智能嵌入业务流程

如今，人工智能和 Machine Learning 盛行，这也是理所应当的。Machine Learning 框架的发展，加上图形处理单元 (GPU) 专用服务器的使用使得各种新功能得以实现，如，自动驾驶。当然，为了训练 Machine Learning 模型，需要提供大量数据（因此，我在上文中对数据湖做了探讨）。各组织已经开始利用这些 AI/ML 功能来获取以前不可能实现的结果，例如，根据视网膜成像，更好地预测健康状况，或者现场预测停机或硬件故障。企业可以让 AWS 分担其繁重任务来增强自身的 AI/ML 组织脉络，这并非科幻小说，而是每天生产运营中的真实情况。

最后，我想强调一点，世界上最好的算法如果不与业务流程相集成，也将毫无价值。通常，获取洞察力或创建数据科学模型是很容易的，但是，要将其集成到您的保险规则引擎或零售平台则不太容易，因为这些系统通常不具备集成外部数据源或 API 的功能。这是一个很好的机会，可以考虑将这些系统迁移到云中，以利用所有可用的服务来帮助您实现系统的现代化或重新构建。

组织洞察力

在公司中构建高级分析能力需要的不仅仅是技术。通常，组织面临的最大难题与组织架构 - 流程、监管和人员有关。要想您组织进行的分析投资取得成功，您需要记住哪些方面？

> "世界上最好的算法如果不与业务流程相集成，也将毫无价值。"

首先建立卓越分析中心

从战略意图到实质性进展，首先需要确定和选择一个领导者和团队来引领变革，并设立卓越分析中心 (COE)。开始时，通常组建小型团队，由一些跨职能部门人员领导，然后将团队发展壮大，以满足日益增大的需求。

许多大型企业已经设立了共享服务组织，以处理商业智能数据或报表。这些组织可以为卓越分析中心提供技术和业务人才。与 IT 基础架构类似，这些组织不仅应该提供人才，而且还应该是这项工作的主要推动者和支持者。因为随着时间的推移，这些报表共享服务组织需要不断发展，以适应或成为卓越分析中心的一部分。初创人员通常包括数据工程师和架构师、商业智能分析师和数据科学家。该团队应该由能够跨多个组织、业务部门和后勤团队（如财务和 IT）开展工作的人领导。

满足您客户的所有需求

各组织需要转变心态，从“你必须使用我们的报表解决方案，你会喜欢它的”转变为“您需要什么样的分析，我们如何帮你实现分析？” 共享服务报表组织通常只是报表推送者，并不能回答员工、企业领导和客户提出的疑难问题。

因此，在设立新的卓越分析中心时，重要的是要为该团队制定原则，通过原则为团队的行动和决策设定预期值。

卓越分析中心需要为两种类型的客户提供服务：

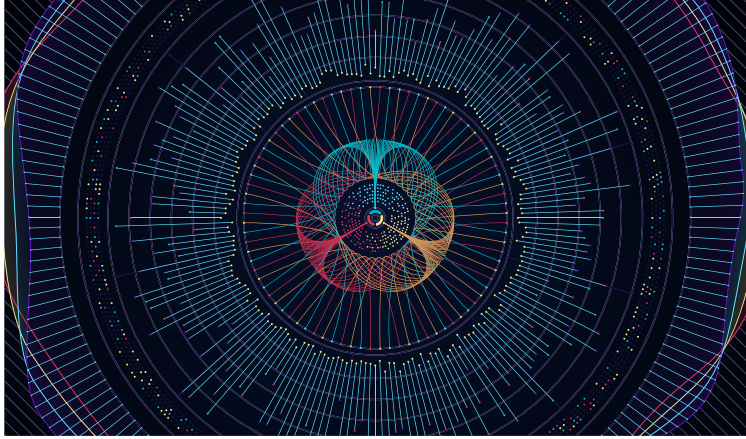
- **数据和分析消费者**： 决策者、数据科学家、商业智能 (BI) 分析师和开发人员。这类客户通常注重快速获取洞察力和访问数据的能力，以及他们用于处理和显示数据的工具和服务的质量。
- **数据提供者**：向平台提供数据的应用程序、基础设施和设备的所有者。这类客户需要这样的服务，即能够轻松将数据发布到分析平台并定义数据合同。这包括数据的域模型、刷新频率和策略定义，例如，概述谁可以访问数据的安全策略。

> “仅仅让您的客户拥有最新工具是不够的，还要让您的客户（生产者或消费者）能够轻松获得他们需要的东西。”

分析功能和平台需要为这两种类型的客户提供服务，如果无法满足他们的需求，则分析工作将没有商业价值。因此，拥有一种机制是非常重要的，它可以从非常庞大、多样化的潜在商业部门和群体中捕捉这两类客户需求。一些组织设立了咨询委员会或与几个关键利益方合作，以促进需求满足。这是一个没有唯一标准答案的问题，但是拥有一个能够捕捉客户声音并对客户需求进行优先排序的机制非常重要。

重新审视卓越中心

卓越分析中心不仅服务于云项目，而且还提供了一套专门满足分析需求的云服务。在过去，报表和 BI 组织常常会提供一种解决方案来满足所有人的需求（一体适用策略）。在这个大数据、丰富可视化、自动化决策、人工智能和 Machine Learning 飞速发展的时代，只有一个技术堆栈是不可能的。仅仅让您的客户拥有最新工具是不够的，还要让您的客户（生产者或消费者）能够轻松获得他们需要的东西。



卓越中心有可能沦为礼宾司服务。卓越中心可以很好的应对某些类型的服务请求，但是，如果没有可扩展的自助服务机制以及透明化的优先级请求和监管流程，卓越中心很快就会因服务请求积压而不堪重负。卓越分析中心需要设计和构建自助、安全、可操作且可扩展的数据平台，以及不断发展的技术生态系统，以处理、分析和提供洞察力。

虽然建成数据驱动型组织并非能够一蹴而就，但找准您面临的数据难题、规划客户服务计划以及让您的团队在正确的时间提供合理的价值，就是您的组织朝着正确方向迈进的一步。

关于作者

Joe Chung，Amazon Web Services的企业策略家与宣传官。