

已获分发许可

# 推动业务变革的十大数据与分析技术发展趋势

发布日期：2019 年 4 月 11 日 - ID：G00379563 - 阅读时间约 65 分钟

作者：Rita Sallam、Donald Feinberg、Mark Beyer、W.Roy Schulte、Alexander Linden、Joseph Unsworth、Svetlana Sicular、Nick Heudecker、Ehtisham Zaidi、Adam Ronthal、Erick Brethenoux、Pieter den Hamer、Alys Woodward



这些数据和分析技术趋势在未来三到五年拥有巨大的颠覆性潜力。数据和分析负责人必须衡量这些趋势的业务影响力，并相应地调整运营、业务和战略模式。

## 概述

### 重要调查结果

- 数据和分析在数字化转型中的扩展和战略性作用正在增加数据的复杂性、需分析变量的数量及实现成功所需的分析类型。这一变化正在突破当前功能和方法的极限。
- 实际上，数据管理、分析内容、应用开发和洞察力共享的各个方面都已开始使用机器学习 (ML) 和人工智能 (AI) 技术，以实现手动任务、分析流程和人类洞察力到行动的自动化执行及增强。

- 支持大规模采用新兴的敏捷数据结构以及可解释的透明洞察力和 AI 的智能功能，是满足新的需求和扩大采用的必要条件。

## 建议

对于您的数据和分析战略而言：

- 向企业领导者介绍您的战略优先事项，以及哪些人工流程可以通过数据和分析实现自动化或得到增强，帮助他们了解相关概念并与之进行互动。
- 建立正式的机制，以识别技术趋势，并对那些会对业务产生最大影响且可纳入到战略和路线图的趋势进行优先级排序。
- 在未来三到五年内采取行动，主动监控、试验或利用关键趋势。当趋势趋于成熟时，不要仅仅被动应对。
- 识别数据、分析和组织功能中妨碍您利用颠覆性趋势的薄弱环节。
- 使用强调学习和奖励创新的成功指标和激励措施。
- 投资非技术趋势，例如数据素养、AI 治理、数据工程、数据故事讲述以及隐私与道德，因为这些都是成功的关键因素。

## 战略规划假设事项

到 2020 年，增强分析将成为客户新购分析和商业智能、数据科学与机器学习平台及嵌入式分析的主要驱动因素。

到 2022 年，通过增加机器学习和自动服务级别管理，数据管理方面的手动任务量将减少 45%。

到 2020 年，50% 的分析查询将会通过搜索、自然语言处理或语音来生成，也可能自动生成。

到 2021 年，自然语言处理和对话分析将会把分析和商业智能的采用率从 35% 的员工提高到 50% 以上，其中包括新的用户类别，尤其是前台工作人员。

到 2022 年，图形处理和图形数据库的采用将以每年 100% 的速度增长，进而持续加速数

据准备并支持更复杂、适应性更强的数据科学。

到 2022 年，在利用 AI 和 ML 技术的新的最终用户解决方案中，将有 75% 的解决方案使用商业平台（而非开源平台）构建。

到 2022 年，来自超大规模云提供商（亚马逊、谷歌和微软）的基于云的 ML 服务将在数据科学平台市场中达到数字临界点，即 20% 的份额。

到 2022 年，用户与应用或设备之间的每次个性化交互都将是自适应式交互。

到 2022 年，定制化数据结构设计将主要作为静态基础结构部署，进而迫使组织进入新的“完成成本”重新设计浪潮，以采用更具动态性的数据网络方法。

到 2023 年，超过 75% 的大型组织将会聘请行为取证、隐私和客户信任方面的人工智能专家，以降低品牌风险和声誉风险。

到 2021 年，大多数获许可的区块链用例将会被账本 DBMS 产品所取代。

到 2022 年，超过一半的关键新业务系统将会采用使用实时情境数据来改善决策的持续智能。

到 2021 年，永久性内存将占内存中计算领域内存 GB 消耗量的 10% 以上。

## 分析

### 是什么在推动十大数据与分析技术趋势？

数据和分析在数字化转型中的扩展和战略性作用正在增加数据的复杂性、需分析变量的数量、分析的类型及实现成功所需的分析速度。随着复杂性的增加，风险和挑战变得更加微妙，其潜在破坏性也更大，例如潜在的偏见以及对分析与 ML 和 AI 模型的透明度和信任的需求。

越来越近乎实时的优化决策所需数据的规模、复杂性和分布式特性意味着，僵化的架构和工具正在逐渐退出。这种复杂性正在挑战现有方法的极限，使得您必须以前所未有的速度在数据和分析领域开展创新，以满足新的要求。

同时，要发挥影响力，数据和分析技术必须覆盖整个企业，并在企业内扩展，甚至扩展至覆

盖客户、合作伙伴和产品本身。

本研究涵盖的战略技术(如图 1 所示)都是您不能忽略的趋势。他们有潜力颠覆您的企业,并将在未来三到五年内加快普及。

许多趋势是相互关联的,因为它们受到许多相同技术颠覆的推动,但会影响到数据和分析技术堆栈的不同部分。

所有趋势都有三个共同的属性 - 它们支持**智能**、**新兴**的数据和分析技术,并且具有**可扩展性**,以便为您提供无处不在且由 AI 和 ML 驱动的洞察力以及以数据为中心的敏捷架构:

- **智能**: 包括 AI 和 ML 技术在内的高级分析是未来平台、解决方案和应用的核心。目前,我们已经看到了这一趋势的萌芽。实际上,数据管理、分析内容和应用开发以及洞察力共享的每个方面几乎都融合了 ML 和 AI 技术。您可以利用它们自动执行或增强手动任务和分析流程,并帮助各种用户角色自动执行或增强人工洞察力到行动的转化。所有数据和分析平台组件中的增强智能将使得相关技能大众化,从而帮助您使用这些功能,并将功能在整个企业中的普及率扩大到前所未有的水平。
- **新兴**: 随着越来越多的数据模型和分析模型由自动生成的模型而不是由代码创建,在 AI 和 ML 的辅助下,数据和分析本质上变得更具归纳性。数据衍生出的结构和洞察力趋于多样化,而不是对数据强加一个统一的结构。支持采用新兴、敏捷和透明的数据基础架构的智能功能是满足新的需求和扩大采用范围的必要条件。数据和分析平台的成功取决于建立信任、责任制、治理和安全的基础,该基础要尊重隐私并提升数字道德。
- **可扩展**: 数字颠覆带来的巨大挑战(数据过多)也创造了前所未有的机遇。大量数据与功能日益强大的云处理功能(用于数据管理和数据科学)以及(十大趋势支持的)新兴功能整合一体。这种情况下,您有机会大规模训练和执行算法,这是发挥 AI 所有潜力的必备要素。要发挥这些潜力,您不仅需要大规模进行处理,还需要更广泛地采用高级分析。

颠覆出现的频率越来越快,这要求数据和分析负责人必须拥有正式的机制来识别技术趋势,并优先考虑对其竞争优势具有最大潜在影响的技术趋势。您应该将这些趋势纳入战略计划,或者如果它们是当前计划的一部分,则您应该根据趋势对业务优先事项的支持程度,重新审

视这些趋势。

数据和分析负责人应主动管理他们监控、试验或部署新兴趋势的方式。实施成功的衡量标准和激励措施，在试验时侧重于学习和奖励创新，这有助于进一步为成功做出贡献。

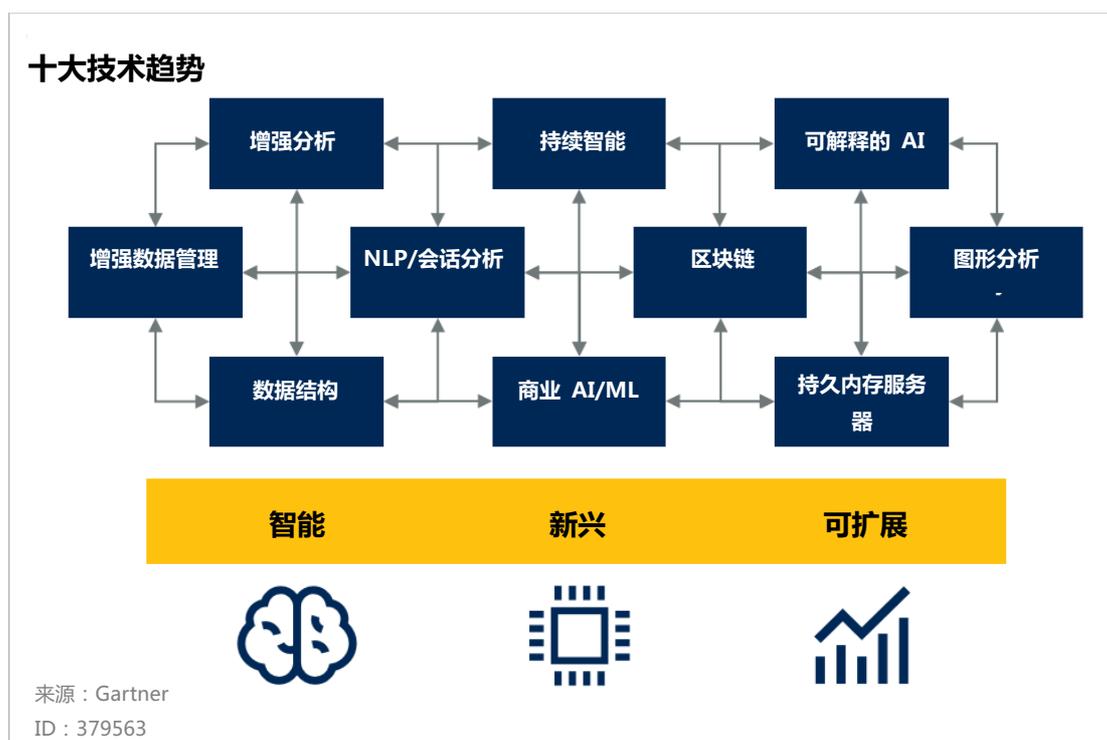
请注意，尽管我们在图 2 中列出了一些与技术无关的关键数据与分析趋势，但本报告并未详细介绍这些趋势。但是，您必须对非技术趋势（例如数据素养、AI 治理、数据工程、数据故事讲述以及隐私与道德）进行同等的投入和投资，这一点很重要，因为这些趋势都是重要的成功驱动力。

此外，本报告也未涵盖成为主流不足三年的当前趋势（例如云商业智能或自助服务）以及成为主流已满五年以上的趋势（例如量子计算）。也就是说，本报告关注的是成为主流的时间在 3 年到 5 年之间的趋势。

如需更深入地了解这些趋势，请查阅 “The Future of Data and Analytics: Tales and Trends From the Center to the Edge” 。

有关当前趋势看法和可行建议的信息，请参阅 “Leadership Vision for 2019: Data and Analytics Leader” 。

图 1：有助于推动业务变革的十大技术趋势



AI = 人工智能; ML = 机器学习; NLP = 自然语言处理

## 趋势 1：增强分析

智能



新兴



可扩展



到 2020 年，增强分析将成为新购买分析与商业智能、数据科学与机器学习平台以及嵌入式分析的主要驱动力。

借助增强分析功能，您可以利用机器学习和 AI 辅助的数据准备、洞察力生成和洞察力说明，帮助业务人员和分析人员更好地探索和分析 BI 与分析平台中的数据。此外，它还能自动开发、管理和部署数据科学和 ML 模型的很多方面，从而增强专业数据科学家和普通数据科学家的实力。

### 为什么立即行动？

- 随着数据变得越来越复杂，业务人员和决策者正在接受数据、风险、复杂性和混乱的洗礼。他们正在艰难地识别什么是最重要的，以及哪些是最佳措施，或者哪些是应该规避的措施。数据集组合规模越来越大，越来越多样化，这意味着，您需要分析、探索、建模、测试和治理的变量与关系越来越多。在不引入偏见的情况下，使用当前的手动方法来完成这些任务变得越来越难，甚至成了不可能完成的任务。
- 在整个分析堆栈中，工具变得更加便捷和敏捷，从而支持更大的访问权限和自助服务。但是，许多流程在很大程度上依然是手动流程，并且容易出现偏见。其中包括管理数据、准备数据进行分析、分析数据、构建数据科学和 ML/AI 模型、解读结果、用数据讲故事以及使洞察力变得切实可行等流程。所有这些活动都有一个基本组成部分，那就是您必须提前知道有关数据关系的假设。如果使用当前的方法，用户不可能探索所有可能的组合和模式，更不用说确定他们的发现是否是所有可能的选项中最相关、最重要和最可行的选项。

- 依靠业务用户找到模式，依靠数据科学家手动构建和管理模型，这可能会导致他们基于自己有偏见的假设探索并证明某些关系。结果就是，他们可能会错过关键的发现，并得出不正确或不完整的结论。这会对决策、行动和结果产生不利影响。

### **该趋势能提供哪些支持？**

- 增强分析功能可（结合用户的情境）自动发现和显示业务中最重要的洞察力或变化，以优化决策。它能够快速做到这一点，且您需要掌握的数据科学和机器学习技能更少，也不必像采用手动方法时那样事先了解数据关系。
- 增强分析使用 ML/AI 技术来自动执行数据科学和 ML/AI 建模的关键方面，例如特征工程、模型选择 (autoML)、模型操作化、模型说明以及最终的模型调优和管理。因此，高技能型数据科学家有更多的时间专注于创造性任务，以及构建和操作最相关的模型。可以测试的模型数量大大增加，并且迭代和测试模型的时间与成本大大减少。
- 利用增强分析创建的许多自动生成的和人工增强的 ML 模型已被嵌入企业应用中，例如人力资源、财务、销售、营销、客户服务、采购和资产管理部的应用。这有助于您优化难以捉摸的数据和分析的“最后一公里”，即，所有员工的决策和行动，而不只是分析人员和数据科学家的决策和行动。
- 作为数据准备、更广泛的数据管理、高级分析和 BI 以及数据科学和 ML 平台的关键功能，增强分析功能正在迅速成为主流。
- 您还可以使用 NLP 和会话式界面（也是十大趋势之一）部署增强分析，让组织中更多的人使用增强分析，并在不使用高级技能的情况下，根据数据做出预测，获得可行的洞察力。它将为缺乏技能或无法访问分析和 BI 平台提出自己问题的人提供深入的洞察力。

### **该趋势如何影响您的组织和技能？**

- 增强分析让所有业务角色都能访问从分析中挖掘的洞察力。尽管这种趋势能帮助您减少对专业分析、数据科学和机器学习技能的依赖，但是它需要整个组织更多地关注数据素养。

- 增强分析减少了对专业数据科学和机器学习技能的需求，可以生成、操作和管理高级分析模型。它还支持“普通数据科学家”（包括必须将 ML/AI 嵌入应用的业务分析师和应用开发人员）创建数据科学和 ML 内容。这样，专业的数据科学家就能提高工作效率和协作能力，也就能投入更多精力去执行高价值的任务。利用增强分析功能部署流程以增强不同角色之间的协作对于成功至关重要。

## 用例

用例覆盖了所有行业和领域，在这些行业和领域中，所分析的数据和变量过于复杂，您用当前的方法已经无法全面、准确地探索这些数据和变量。

部分示例如下：

- **银行业**：在实施增强分析之前，银行有针对性地为年长客户提供财富管理服务。借助增强分析，他们发现实际上年轻的客户（年龄在 20 至 35 岁之间）更有可能接受财富管理服务。
- **农业**：在实施增强分析之前，数据科学家花费了数月的时间来建立模型，才能从成千上万的种子中找到最好的杂交种子组合卖给农民。借助增强分析，专业的领域遗传学家接管了整个流程，并将流程周期缩短至几天。
- **医疗保健行业**：在实施增强分析之前，美国医疗保险公司将患者的疾病评估结果作为运输（救护车）成本的主要驱动因素加以跟踪。借助增强分析，他们发现主要的成本驱动因素是 12 岁以下的孩子。调查发现，整个路程是按人头收费，其中包括陪伴患病儿童的父母。

## 建议

- 通过试运行增强分析来解决目前需要用耗时的手动分析才能解决的高价值业务问题，探索补充现有数据和分析计划的机会。
- 通过加强专业数据科学家与普通数据科学家的协作，回测和证明价值，建立对机器辅助模型的信任。了解机器辅助模型的局限性，相比尖端技术，机器辅助模型能够与经验证的算法实现最佳协同。
- 监控老牌数据和分析提供商、企业应用供应商和初创企业的增强分析功能与路线图。

- 评估前期设置、数据准备、模型的开放性和可解释性，支持的变量数，提供的算法范围以及模型的准确性。
- 制定战略以培养角色、职责和技能，并增加对数据素养的投资。

### 相关报告推荐

“Pursue Citizen Data Science to Expand Analytics Use Cases”

“Augmented Analytics Is the Future of Data and Analytics”

“How Citizen Data Science Can Maximize Self-Service Analytics and Extend Data Science”

“Rebalance Your Integration Effort With a Mix of Human and Artificial Intelligence”

### 趋势 2：增强数据管理

智能



新兴



可扩展



到 2022 年，通过增加机器学习和自动服务级别管理，数据管理方面的手动任务量将减少 45%。

供应商正在添加 ML 功能和 AI 引擎，让自配置和自调优流程变得无处不在，就好像 ML 和 AI 功能正在各个数据管理类别实现分析、商业智能和数据科学革新一样。这些流程能自动执行很多手动任务，使技术技能较低的用户在使用数据时拥有更多自主权。这样，高技能型技术专家就能专注于更高价值的任务。这一趋势正在影响所有企业数据管理类别，包括数据质量、元数据管理、主数据管理、数据集成和数据库。

### 为什么立即行动？

数据量日益增长，数据源越来越多样化。与此同时，数据的获取和生成又是源自可信和不可信数据源的组合。由于技术技能的短缺，为响应治理策略、规则和流程而引发的自动化数据

管理任务需求正在呈几何级数增长。在分布式环境中，即使以前单调（并且容易完成）的任务也太多了，无法满足以数字业务的速度扩展的关键需求。

随着客户将数据资产迁移到云端，您将跨越标准化配置库获得更多使用情况统计信息。随着每个提供商对识别数据管理模式越来越有信心，构建允许云提供商使用和执行自动化数据管理战略的模型变得可能。

凭借这些模式以及识别每种方法的准确性的能力，您可以更轻松地“克隆”最有效的方法和部署选项。例如，“我需要另一个这种配置的沙箱”，那么只需单击几下即可（甚至只需机器发出指令）。结果就是大大提高了部署敏捷性。

### **该趋势能提供哪些支持？**

- 您可以从数据利用率、用户和用例中推断出新的元数据，而不是使用通常不再与实际数据捕获/写入和后续使用同步的描述性元数据。
- 通过执行适当的流程，使用盘点和编目并自动发现对数据管理至关重要的语义、分类和本体，您可以积极运用所有元数据类型，构建成熟的数据结构。组织需要轻松地知道他们拥有的数据、含义、传递价值的方式以及数据是否可信。通过利用现有系统的统计信息及其可用容量、已知资源、使用每种资源的成本参数，以及基础架构组件之间的通信能力，策略级指令将确定在何处执行数据操作。他们可以管理生产部署，甚至可以在必要时重新配置部署。
- 需要一致的性能和持久的操作时，您将以更加定制化的方式，在新环境中部署由微服务、API、数据管道和数据质量/治理服务组成的数据结构方法。
- 数据融合功能可以跟踪用例使用了哪些资产，并生成知识和利用率图。当遇到新的数据资产时，融合引擎将分析这种数据资产与其他常见数据资产的相似性。它们将确定新数字资产与现有数据/用例的关联性，然后通知其他自动化系统有新的数据可用，并且可以将这些数据纳入自己的系统。
- 动态数据识别允许用“流式”方法评估新的和现有的数据资产，并使用关于数据资产的累积信息来开发相关的事件模型。随着时间的推移，用例将形成处理需求，指出需要为操作和/或分析用例提供哪些数据，以及提供数据的速率或频率。

### **该趋势如何影响您的组织和技能？**

到 2023 年,数据管理系统中支持人工智能的自动化技术将使企业对 IT 专家的需求减少 20%。

对数据和分析组织技能的影响：

- 增强数据工程师的能力，或使某些数据工程任务自动化。
- 提醒数据工程师注意潜在的错误、问题和对数据的其他解读。
- 创建对错误、问题和对数据的其他解读的自动系统响应。
- 增强使用公开可用数据、合作伙伴数据、开放数据和其他资产的能力，这些资产目前很难确定是否适合使用。
- 自动化数据查询，从而模拟数据发现，甚至评估新资产是否符合已知或现有模型的“可信度”。

对分布式数据管理的影响：

- 持续监控数据管理环境的容量和利用率，以重新分配资源计划，甚至实现跨越多云环境和本地实施的分配。
- 执行优化和性能管理，不再需要手动（或人工）决定何时创建数据的中间、临时或永久副本以提高操作或分析性能。
- 基于策略的决策引擎将法规需求捕获为元数据配置，然后管理热/暖数据用例、冷数据使用情况、归档期望，甚至数据清除或数据粉碎，以维护法律和审计合规性。

## 用例

增强数据管理功能将机器学习和人工智能技术用于以下方面：

- **数据质量**：扩展不同数据源中主数据的分析、清理、链接、识别和语义协调。
- **主数据管理 (MDM)**：用于 ML 驱动的配置以及记录匹配和合并算法的优化。
- **数据集成**：通过推荐甚至自动执行重复的集成流程，简化集成开发流程。
- **数据库管理系统 (DBMS)**：用于存储、索引、分区、数据库调优、补丁、升级、安全性和配置的自动化管理。

- **元数据管理**：除了数据沿袭之外，ML 还可以用于评估数据规则、填充语义框架，以及帮助您跨越日益多样化的数据源发现和摄取元数据。

## 建议

- 计划减少在平衡物理基础架构和逻辑数据管理需求方面具有专业知识的技能型人员和员工。从研究包含动态硬件/基础架构配置的容量规划工具入手。
- 制定相关战略，隔离多云和本地场景所需的数据资产，用于潜在的复制/副本实例，同时考虑其物理位置的法规和隐私限制。
- 在组织中培养数据素养。虽然自动化技术将减少使用数据的技能障碍，但它仍然需要用户理解数据并正确使用数据。这将需要您进一步扩展元数据的访问和利用。

## 相关报告推荐

“Toolkit: Map Your Data Management Landscape With the Data and Analytics Infrastructure Model”

“Toolkit: Comparative Total Cost of Ownership Calculator for Cloud and On-Premises DBMS Deployments”

“Data Hubs, Data Lakes and Data Warehouses: Choosing the Core of Your Digital Platform”

“Data Management Strategies: Navigating Diverse Roles, Use Cases and Markets”

## 趋势 3：NLP 和会话分析

智能



新兴



可扩展



到 2020 年，50% 的分析查询将会通过搜索、自然语言处理或语音来生成，也可能自动生成。

到 2021 年，自然语言处理和对话分析将会把分析和商业智能的采用率从 35% 的员工提高到 50% 以上，其中包括新的用户类别，尤其是前台工作人员。

NLP 为任何用户提供了一种更简单的方法来询问有关数据的问题，并获得对呈现的洞察力的解读。就像谷歌这样的搜索界面让普通用户和消费者都可以访问互联网一样，搜索和自然语言查询 (NLQ) 方法能够提出有关数据的问题，让普通业务用户也能访问数据和分析功能。其中包括那些没有技能或无法使用当前基于可视化的点击式分析和 BI 系统的用户。

目前，供应商采取了各种各样的方法来支持这样的界面。有些供应商使用简单的关键字匹配（与谷歌的“搜索”风格类似）。还有一些供应商使用强大的自然语言查询，且支持用户使用行业特定的语言执行更复杂的查询。例如，早期的模型可能允许用户请求“各个产品的销售情况”，以生成基本条形图。更强大的模型将支持很多问题，比如“向我展示在方圆 50 英里最畅销的产品，并将今年与去年进行比较”。尽管“最畅销”不是一个关键字，但该平台知道，它必须应用一个排名函数，执行地理分析，然后执行一个子查询来进行逐年比较。

结合利用自然语言查询 (NLQ) 与自然语言生成 (NLG)，从而利用自动生成的文本解释和解读结果。文本可能描述屏幕上的可视化内容，也可能描述没有专门请求的数据中的相关模式，并使用自然语言提供操作建议。

目前，糟糕的数据素养阻碍了分析和 BI 计划发挥影响力。通过改进数据的一致性解读，不管用户技能如何，NLG（单独使用或与 NLQ 和增强分析相结合）都有可能提高整个组织的数据素养水平。在这里，文本描述也可以采用书面文本形式或语音生成的形式，或者两者兼而有之。不同的产品有不同的语言控制能力，如控制唠叨和语调的能力，并且文本中基于表格的文本、模板驱动的文本或自动生成的文本的比例也各有不同。

会话分析仍然属于新兴技术，但它允许口头提出此类问题，使 NLQ 和 NLG 的概念更进了一步。您可以通过数字助理（如 Amazon Alexa）、手机或其他设备开展这种语音交互。除了支持语音查询之外，这些功能正以虚拟人工智能助手或机器人的形式支持会话。NLP/会话式界面也将利用增强分析所生成的洞察力和嵌入式 ML/AI 模型。虚拟助理可能会回答最初提出的问题，但随后会利用借助增强分析自动生成的洞察力以及用自然语言提出的建议加以详细阐述，比如，“紫色产品的销量在第 30 周出现激增。您想知道发生了什么以及背后的原因吗？”此外，它还将作出预测并规定行动。

## 为什么立即行动？

随着组织转型成数字企业，分析成为了一个关键驱动因素。让所有员工都能接触到分析功能生成的洞察力，这将是推动企业变革的关键所在。然而，基本上只有掌握不同级别分析和专业技能的高级用户、业务分析人员和专业数据科学家能够访问 BI 和分析以及数据科学平台生成的分析内容。

尽管分析和 BI 平台的易用性有了很大的改进，但是它们的采用率仍然很低，因为它们专注于支持高级用户而非主流业务用户。

早期基于搜索的 BI 平台需要大量管理设置，只能解决简单的问题，并且仅能处理小型数据集。最新的解决方案能解决更复杂的问题，并且几乎不需要高级配置，就能实现自动化索引，并快速处理大型数据集。

随着大规模并行处理技术的问世，通常基于知识图的复杂数据模型（包括结构化和非结构化数据）扩大了使用范围，而且 GPU 和云平台开始用于处理数据，这一切都使规模化 NLP 成为可能，并且有望实现更广泛的普及。

借助基于语音的数字助理以及自由访问数字助理的开放式 API，供应商无需自行开发技术，即可用语音增强他们的平台。聊天机器人的情况也一样。

## 该趋势能提供哪些支持？

- 目前，大多数分析和 BI 工具都要求用户选择数据元素并将其放在页面上，以创建查询和可视化分析。NLP/会话分析将易用性提升到了一个新的水平，并将查询变得像谷歌搜索或会话数字助理（如 Alexa）一样简单。
- 任何用户都可以使用文本或语音来提问，并且问题和答复变得越来越复杂。NLP 逐渐成为了一个界面，让您可以查询增强分析自动生成的洞察力，并与这些洞察力交互。
- 通过结合利用 NLP 与增强分析（包括自动化洞察力生成），用户能够快速地“大海捞针”，并利用会话分析和自然语言生成技术，获得最重要且可行的洞察力。
- 有些更强大的 NLP 界面还拥有行业或领域特定的分类和语言（如人力资源、财务、医疗保健或金融服务），以便更准确地解读短语。

- NLP 和会话式界面被嵌入了分析和 BI 平台、数字个人助理、聊天机器人和应用中。
- NLP 提供了直观的形式供人与系统进行交流。NLP 采用计算语言技术（包括传统技术和机器学习），旨在识别、分析、解读和生成自然语言。
- NLP 和基础知识图正在扩展，以分析非结构化数据和其他数据类型，并充当数据科学以及 ML/AI 模型的基础。

### **该趋势如何影响您的组织和技能？**

会话分析可以大大提高每个员工的分析采用率，而不是主要由高级用户和业务分析人员使用分析功能，从而提高业务影响力。然而，随着企业内拥有不同技能水平的员工都能获取更强大的洞察力，您有必要正式在企业范围内重点关注如何提高所有用户的数据素养。

### **用例**

任何用户（从首席级高管到分析人员、再到运营人员）都可以通过语音与虚拟个人助理或他们的手机进行交互，并要求执行与自己相关的分析；他们不再需要登录复杂的仪表盘。

比如，通过结合利用增强分析功能，销售人员可能会要求对销售或管道进行分析。或者系统可能已经了解到销售经理也会查看这些信息。根据该人的角色和/或行为，系统将以文字或语音的形式向他们提供对统计上重要的变化驱动因素的解释或叙述。然后，系统可以（通过设备）发送可视化内容，以根据用户的角色展示重要的趋势、模式或异常值。预测和说明性建议也可以传达给销售经理。会话分析也将嵌入到每个员工使用的的应用工作流中。一开始，用户可能会分析增强分析技术自动生成的可视化内容或洞察力，任何用户都可以用文本或语音形式提出问题，进一步探索。

### **建议**

- 评估您的分析和 BI 平台与企业应用供应商（以及那些初创公司）的能力、路线图和合作伙伴关系。
- 评估解决方案的成熟度和可扩展性，特别是解决方案的集成和易用性、预先设置/配置需求以及支持的语言和分析类型。
- 投资数据素养是一个关键成功因素。

### **相关报告推荐**

## 趋势 4：图形分析



到 2022 年，图形处理和图形数据库的采用将以每年 100% 的速度增长，进而持续加速数据准备并支持更复杂、适应性更强的数据科学。

图形分析是一组分析技术，让您能够探索组织、人员和事务等感兴趣的实体之间的关系。图形分析的一个应用（即，支持图的语义知识图）为许多 NLP/会话式界面和数据结构奠定了基础，充实并加速了数据准备和数据科学。

### 为什么立即行动？

业务人员针对结构化和非结构化数据，提出了更复杂的问题。为了执行复杂的分析，通常您需要混合来自多个数据源和多个业务部门的数据以及越来越多的外部数据。当要确保内容和媒体的相关性时，天气、经济、生产、人员、法规、绩效指标、备选资产与主要资产的对比等概念，甚至时区变化等基本概念都非常复杂。使用传统查询工具或查询语言（大多数工具或平台最常用的语言是 SQL）大规模分析这些内容是不切实际的，或者说只有在某些情况下才可以实现这种分析。

许多用例将图形当作支持技术使用。企业通常通过可视化内容呈现图形分析结果，供业务用户使用。在图形可视化内容中，您可以将一个指标拖到另一个指标中，或删除指标的某个部分，然后将该指标与其他数据集重新组合。图形分析由确定跨数据点“连接性”的模型组成。

经过改进后，可扩展且低成本的处理选项（如云和 GPU）使得图形分析和数据库成为了企业加速采用的首选技术。

### 该趋势能提供哪些支持？

图形分析展示了几个趋势或数据点之间的紧密联系。例如，一个自动照明系统可以设置为在

黄昏时开启,但它也可能在恶劣天气情况下开启。系统可能结合利用了时段和季节变化与“清醒时间”评估模型。所有这些输入有可能相互关联,也可能相互不关联,而图形分析(根据它们的关联程度)将这些输入显示为紧凑的数据“节点”。它还揭示了这些节点是紧密压缩的节点还是只是松散的连接。当条件变化时,节点之间的分界点可能也会变化,并且可能会动态地出现一组全新的节点。此外,图形分析还显示了某件事是否仅仅是一种相关性,或者这两件(或更多)事是否有依赖关系,一件事是另一件事的前提,或者它们只是单纯的“因果”关系。这意味着,节点被显式或隐式连接,表明影响级别、交互频率或概率。

图形模型确定数据点之间的“连接性”,并根据影响级别、交互频率和概率创建集群。一旦开发和训练了高度复杂的模型,您就能更轻松地存储输出,这要归功于扩展的功能、计算能力和采用的图形数据库。用户可以直接与图形元素交互以发现洞察力,并存储分析结果和输出,在图形数据库中重复使用。

因此,图形数据库为图形模型的存储、操作和分析提供了一个理想的框架,尽管某些图形分析供应商有自己的图形分析引擎,不需要单独的图形数据库。

通过生成有关感兴趣的不同实体(人、地方和事物)如何关联的动态图,而不是采用更静态的关系模式,您可以提供更接近人类知识表示的更深刻见解。例如,它可以轻松地结合和关联运动应用、饮食计划工具、病历与医疗新闻动态中的数据,并从中挖掘动态的洞察力。

图形技术为创建更丰富的语义图或知识库提供了基础,这些语义图或知识库可以改进增强分析模型并充实会话分析。此外,图形分析还支持您创建新兴元数据管理和数据目录。它通过捕获以下方面的知识来做到这一点:您拥有什么数据?数据存储在哪里?数据之间有何关联?谁在何时出于什么目的以何种方式使用数据?ML模型可以利用这种洞察力提出建议,为企业及其应用提供更加个性化、自动化和妥善管控的洞察力。

### **该趋势如何影响您的组织和技能?**

- 您需要新的技能。其中包括图形特定的标准;图形数据库;各种技术和语言,比如资源描述框架(RDF)、SPARQL协议和RDF查询语言(SPARQL);以及新兴语言(如Apache TinkerPop或最新的开源Cypher)。这些技能的缺乏将阻碍您采用图表。
- 图形分析的商业化依然处于早期阶段,仅有少数新兴企业提供了商业化的图形分析。

- 虽然目前的技术仍然需要您掌握专业技能,但市场上已经出现了 SQL-to-graph 解释器。这些解释器将基于图形的程序转换成分解的程序 SQL (也可以再转换回来)。这些功能有助于支持图形技术与现有数据集的兼容。

## 用例

用例的数量随着复杂分析的需要而增加。从欺诈检测到客户影响者网络,再到社交网络和语义知识图,用例范围很广。会话分析、健康顾问、金融犯罪和风险检测流程也可以利用图形功能。

其他专业应用包括：

- 在路点之间进行路线优化,以改进运输、配送甚至人流量。
- 开展市场购物篮分析,表明存在巧合或依赖关系的产品。
- 检测欺诈,以识别围绕互联网的“攻击者”群体的活动集群。
- 通过分析社会网络,确定谁是影响者、“金丝雀”、决策者和阻碍者。
- 优化 CRM,以确定“下一个最佳产品”这类任务的成功概率。
- 不确定位置时,利用位置情报,根据已知数据点确定可能的路线或位置。
- 任何类型的网络系统中的负载均衡,如通信或实用程序,但也有计算机网络这类更简单的东西。
- 特殊形式的劳动力分析,如企业社交图谱和数字工作场所图形。
- 分析对象、资产和条件相关网络的最近交易时间、交易频率和交易金额,可以最佳利用率指定最优资源。
- 执法调查,以隔离遗漏或未知的风险甚至身份(如儿童伤害案件中的第三方关系)。
- 流行病学,用于分析交叉的环境、病人的整体状况、饮食、运动、急/慢性疾病分析,以及治疗和化学/药物方案的相互作用。
- 基因组研究,用于分析基因的相互作用,以确定潜在的现有遗传疾病病情及其治疗方案,从而辅助有针对性的医疗实践。

- 检测洗钱，以揭示“犯罪者”之间的关系，从而识别恶性和良性行为。

## 建议

数据和分析负责人应该：

- 评估将图形分析纳入分析应用和战略的机会，同时用于分析应用，并以语义指示图的形式成为基础数据结构的一部分。
- 评估传统供应商和新兴供应商提供的产品，但要认识到大多数解决方案都趋于成熟，并且通常专注于特定的领域和垂直行业。
- 探索新的 graph-to-SQL 解释器的使用，但也要投资开发必要的独特技能和能力。

## 相关报告推荐

“An Introduction to Graph Data Stores and Applicable Use Cases”

“The Future of Data and Analytics:Tales and Trends From the Center to the Edge”

“Making Big Data Normal With Graph Analysis for the Masses”

## 趋势 5：商业（而非开源）AI/ML 将主导市场

智能

可扩展



到 2022 年，在利用 AI 和 ML 技术的新的最终用户解决方案中，将有 75% 的解决方案使用商业平台（而非开源平台）构建

### 为什么立即行动？

目前开源平台（包括 Python、R、Apache Spark、H2O.ai、Anaconda 和 TensorFlow）占据主导地位，这标志着市场的不成熟。在过去的五年里，算法和开发环境中持续快速创

新主要发生在开源选项中。然而，商业供应商现在正在将连接器融入开源生态系统中。重要的是，他们还提供了扩展 AI 和 ML 所需的企业功能，而这是开源技术所欠缺的。这些功能包括项目和模型管理、复用、透明度、数据沿袭、平台内聚性和集成。

IT 巨头们对开源创新反应迟缓，他们直到 2017 年才开始提供可行的产品。2018 年，Amazon 宣布推出现有的 ML 平台 (SageMaker)；Gartner 预测，谷歌极有可能会选择把它的大部分 ML 创新成果放在 Kubernetes 框架中。多年来，IBM 和 SAS 一直在弱化其传统的 SPSS 和 Enterprise Miner 平台，目前他们聚焦的是更现代的平台 (IBM Watson、SAS Visual Data Mining and Machine Learning)。SAP 和 Oracle 也对他们当前的 ML 战略和产品实施了重大改进。

到 2022 年，来自超大规模云提供商（亚马逊、谷歌和微软）的基于云的 ML 服务将在数据科学平台市场中达到数字临界点，即 20% 的份额。

开源库和开发环境提供了亟需的民主化数据科学和机器学习，以及创新和灵活性。但是，与过去的预测分析技术一样，灵活性和强大的功能可能会牺牲严格的操作化机制。企业已经创建了许多模型，但是由于尚未大规模投入生产，因此企业尚未完全实现其业务价值。

商用平台提供了更严格、更规范的方法和功能，包括：

- 大数据、数据科学和人工智能领域的快速创新已经开始降温。随着商业供应商将连接器融入开源生态系统中，用户发现他们不仅可以拥有酷炫的创新型开源组件，还可以拥有越来越适合企业使用的商业 ML/AI 平台。
- 开源市场中的创新从来都不是平均分布的。大部分创新都是围绕算法和面向开发者的机器学习平台展开。协作、数据和用户访问权限、操作化、元数据/模型管理以及无缝用户体验领域的创新则要少得多（特别是针对普通数据科学家这样的低技能型最终用户的创新）。

到 2022 年，用户与应用或设备之间的每次个性化交互都将是自适应式交互。

现在，数据科学和机器学习团队开始将业务结果而非生产指标（例如，产出的模型数量，或已启动的项目数量）作为衡量标准。因此，商用平台所需的规范方法逐渐成为了实现业务价值和数据科学团队可持续发展的必要条件。

**该趋势能提供哪些支持？**

- **更高的生产力**：不同操作系统工具的组装和改造需要大量的技能和劳动力，这已经成为商业提供商关注的焦点，同时也为实现商业价值提供了更清晰的路径。
- **AI 民主化**：越来越多的商业提供商将通过精心编排用户体验和“实现互联”，“抚平”通常与开源项目相关的粗糙边缘。通过编排，他们能提供更加集成的功能（参见“Embracing Competition to Evolve and Enable Data and Analytics Product Offerings”）。这也使得大多数企业中更广泛的技能型员工都能开发最先进的 AI/ML。
- **更出色的 AI/ML 规划和路线图**：当前的人工智能战略充满了不确定性、冲突和模糊性。强大的商用平台的回归将通过向 IT 软件基础架构提供具体的锚点，来增强可规划性和 IT 路线图。

#### 该趋势如何影响您的组织和技能？

- 通过提高商用数据科学和机器学习平台的利用率，您能够缩小目前这些领域的技能差距。此外，您还能增加具有不同技能水平的 AI/ML 开发人员之间的协作。

#### 用例

- 来自 Amazon SageMaker、Google AI Hub、Microsoft、KNIME、RapidMiner 的算法和预训练模型市场将成为商用数据科学和 ML 平台的加速器。这将进一步减少围绕收集数据、管理数据、监管数据和以敏捷方式访问数据来构建核心功能的需求。

#### 建议

- 开始或继续提升现有专业人员（例如，业务和技术部门的专业人员）的技能，将他们培养成普通数据科学家，同时逐步迁移到商业生产环境。
- 重视易于使用的增强数据科学/ML 解决方案和即将问世的市场，作为简化 ML 采用的催化剂，或引导 ML 的工作。
- 随着您的数据科学、ML 和 AI 功能不断成熟和扩展，为您部署到生产环境的模型以及模型的业务影响（而不是创建的模型数量）设置成功指标。
- 创建开源审计流程，旨在将开源模型集成到商业生产环境之前先验证开源模型。

- 关注数据管理能力，因为随着算法逐渐商品化，数据将成为 AI/ML 成功的关键决定因素。

### 相关报告推荐

“Embracing Competition to Evolve and Enable Data and Analytics Product Offerings”

“How to Operationalize Machine Learning and Data Science Projects”

“Six Pitfalls to Avoid When Executing Data Science and Machine Learning Projects”

### 趋势 6：数据结构

智能	新兴	可扩展
		

到 2022 年，定制化数据结构设计将主要作为静态基础结构部署，进而迫使组织进入新的“完成成本”重新设计浪潮，以采用更具动态性的数据网络方法。

要从分析投资中获得价值，您必须拥有一个敏捷且可信的数据结构。数据结构通常采用定制的设计，通过以编排方式组合数据集成方法（批量/批处理、消息队列、虚拟化、流、事件、复制或同步），提供可复用的数据服务、管道、语义层或 API。您可以通过添加动态模式识别、甚至基于成本的优化方法（以及其他增强型数据管理功能）来改进数据结构。随着数据结构变得越来越动态，甚至引入了 ML 功能，数据结构也演变成了数据网格网络。

### 为什么立即行动？

与以往任何时候相比，组织成功地部署了更多更大、更复杂的数据和分析实施项目。数据管理和分析市场的稳步发展证明了一致性、语义协调和受治理的信息资产的价值。

组织生态系统中，对业务结果至关重要的资产的动态识别越来越普遍。数据管理技术的融合将为新的和差异化的托管数据服务交换提供通用的平台。背后的驱动因素是您需要对日益增

长和多样化的信息资产来源进行消费、建模和有效可视化。您需要以一致且语义协调的方式执行所有这些活动，并且应该通过活动元数据利用层来启动这些活动。

### **该趋势能提供哪些支持？**

理解数据集成战略交集的最佳方法是从传统实践入手，然后同时考虑架构和基础架构设计中的新实践或新兴实践。这样，您就能将这些实践与组织的理想设计目标进行比较和对比。

现在（以及将来）的数据集成方法必须有效地解决以下三个问题：

1. 元数据推动组织中任何数据资产的整体理解和性能优化。
2. 当涉及到边缘定期联网的设备或传感器时，永久连接的可靠性还没有超过 80%，并且缺乏可靠通信所需的物理基础架构。
3. 处理始终是一个移动考虑事项，并且作为逻辑需求，处理可以进行重新布置，从而将流程引入数据处或将数据引入流程中。这会导致您在选择部署平台时，否定大多数考虑因素（包括云、本地、多云和混合，以及其他组合）。

当前和传统的实践都侧重于结合语义解释的物理整合形式：

- 数据仓库、数据湖和运营数据存储依然是主要的数据存储库，它们代表了特定持久性和性能期望。
- 更传统的数据集成方法仍然倾向于识别和设计特定的目标。出于理解和性能方面的原因，数据仓库、基础运营数据存储系统、甚至逻辑数据仓库都结合利用了物理数据湖、物理数据仓库以及某种类型的统一语义访问层。
- 数据中心代表了传统设计的下一步。它们在地理或域“区域”的分布式数据管理理念中结合使用了数据服务末端，后者通过微服务、API 或桌面即服务 (DaaS) 形式的 P2P 数据服务来确定执行数据交换时采用的倾向性。

如果将元数据、物理管理和处理设计视为部署的主要“极点”，则基础架构设计有两种互补的方法：

- **数据结构**更多的是一种设计方法，主要关注线程两端的用例和位置。线程可能会在中间交叉并进行切换，甚至复用它们的组件，但它们不是动态构建的。它们只是高度可复用的规范化服务。

- **数据网格**是一种完全由元数据驱动的方法。以元数据积累形式提供的统计数据始终包含以下相关信息：数据访问速率；平台、用户和用例访问；系统的物理容量；以及基础架构组件的使用。其他数据点包括基础架构的可靠性、按领域和用例划分的数据使用趋势，以及数据的资格验证、充实和完整性（包括声明的和隐含的信息）。

### 该趋势如何影响您的组织和技能？

特定角色将面临新的工作流程，他们的职责和任务也会有明确的变化。

- **数据工程师**：该角色的职责包括确定数据的可访问性、资格验证、交付和处理。他需要理解如何在广泛的业务流程中使用数据、使用相同数据概念的各种语义方法和基于模式的方法，以及在整合数据时所呈现的各种不同的数据质量。AI/ML 流程将增强数据工程师的能力，这些流程能够识别数据重构、建模、模式生成和数据质量识别的几乎所有初始痛点。这些流程可能提供有关基础架构决策的专家建议，甚至提供实际、动态的资源分配和供应。
- **数据科学家**：该角色将受益于数据融合输出，这些输出将创建有关扩展数据资产的警报。这些警报将专门针对当前的项目进行调整；但是，它们也可以开始识别给定科学家在数据模式方面使用的数据。
- **数据建模师**：数据建模师、数据集成开发人员和负责数据建模的数据库管理员承担的建模工作会越来越，验证工作会越来越多。数据模型总是存在的，有些是正式模型，有些则“不太正式”。在非结构化数据中，模型可以是协作式模型。
- **信息架构师**：信息架构师负责确保信息收集和管理的预期功能与获取、管理、访问和利用数据所需的适当形式保持一致。处理数据结构的信息架构师需要专注于确定数据资产所需的功能并将其作为元数据输入。

### 用例

- **动态数据工程**：更动态、可复用且优化的数据工程管道可以调查和采用数据结构架构。企业只需要设计一次，就可以用“自动”执行模式运行这些工作，其中数据工程师（而不是集成数据流的开发人员）担任裁决者和辅助人员。他们需要灵活地适应不断变化的数据环境（云、多云和混合云），甚至包括与这些数据相关的用户和用例。

- **受控/可信的数据科学**：现在，数据科学家需要端到端的沿袭，并理解数据模型和算法，以提高效率与合规性，而数据结构设计能帮助他们实现数据融合输出，从而提醒他们扩展数据资产。这样，他们就能更清楚地了解自己的活动元数据（包括性能优化、数据质量、设计和沿袭等），并更好地控制自己的项目。
- **逻辑数据仓库架构**：数据结构和网格概念的问世，使得数据管理概念不再包括将数据限制在结合静态语义解释的物理整合中。这限制了集成式数据在异构应用中的复用，这些异构应用需要特定的数据模型和语义解释。

## 建议

- 开始根据所有用例的特征，跟踪数据的来源和起源以及报告和分析的类型。
- 当主要的数据管理和集成方法专注于*连接*（而不是*收集*）数据时，引入数据结构设计。许多用于构建数据结构的工具也被称为“语义工具”，这些工具还可以将数据作为写出缓存层移动到集合中。
- 采用三层的数据结构设计，该设计将行业级数据隔离为统一的设计和通用的概念模型，将组织中普遍存在的（但不是通用的）数据隔离为第二层，将事务值隔离为几乎“仅附加数据”的第三层。
- 构建一个试点数据资产资源利用引擎，该引擎使用支持机器学习的数据目录。首先收集并分析所需的元数据，以确定一个简单的“警报”功能的规范，该功能告诉用户有新的数据可用 - 可能新数据来自数据湖。

## 相关报告推荐

- “Maverick\* Research: Revolutionizing Data Management and Integration With Data Mesh Networks”
- “Toolkit: Identify Data Management Issues Before Moving Data to or From the Cloud”
- “Critical Capabilities for Data Integration Tools”
- “Magic Quadrant for Data Integration Tools”

## 趋势 7：可解释的 AI

智能



新兴



可扩展



到 2023 年，超过 75% 的大型组织将会聘请行为取证、隐私和客户信任方面的人工智能专家，以降低品牌风险和声誉风险。

随着我们转而采用增强分析（包括自动生成的洞察力和模型），这些洞察力和模型的可解释性对于信任、法律合规性和品牌声誉管理至关重要。可解释的 AI 是一组功能，它描述了一个模型，重点介绍模型的优点和缺点，预测模型可能的行为，并识别任何潜在的偏见。它能够清晰地表达描述性、预测性或规范性模型的决策，以确保算法决策的准确性、公平性、可说明性、稳定性和透明度。

AI 治理是为运用 AI、预测模型和算法而分配并保证组织责任、决策权、风险、策略和投资决策的过程。可解释的 AI 为支持 AI 治理提供了技术基础。

### 为什么立即行动？

增强分析自动生成的洞察力和模型的普及，以及复杂的黑盒机器学习和人工智能解决方案的构建，再加上人工智能中不可避免的偏见都推动了可解释 AI 的崛起。

黑盒方法会阻碍人们信任和接受 AI。不透明的算法（如 DNN）将很多高度可变且会很难解释的隐式交互纳入到预测中。

这种方法还可能导致投资被挪用。您可能会得出这样一个结论：您需要投资 AI，以改善或自动执行决策流程的背景开发工作，而实际上，扩大分析的使用以选择决策选项更有价值。通过了解决策是如何制定和建模的，您可以确定最需要哪些技术。

虽然增强型数据科学和 ML 解决了定制 AI 解决方案的技能短缺和开发时间较长的问题，帮助数据科学世界提高了生产力，但是随着采用率的提高，市场对可解释性的需求日益凸显。

增强型数据科学和机器学习能够自动执行算法选择、功能选择和超参数调优等任务。配备可

解释 AI 功能的增强分析解决方案不仅能向数据科学家显示模型的输入和输出 ,还能解释为什么系统选择特定的模型 , 以及增强型数据科学和 ML 所采用的技术。

AI 中的偏见引发了人们对责任和公平的担忧。因此 ,AI 社区和企业负责人关心的是如何发现和解释可能危及社会和企业的偏见影响。例如 , AI 中的偏见会导致政治观点两极分化 , 持续存在的信仰怀疑 , 以及商机之间的虚假联系。

如果没有可接受的解释 , 自动生成的洞察力和模型 , 或与 AI 偏见相结合的黑盒 , 可能会导致人们对监管、声誉、问责制和公平性的担忧 , 并导致人们不信任人工智能解决方案。

### **该趋势能提供哪些支持 ?**

可解释的 AI 通过提高 AI 解决方案和结果的透明度与可信度 , 提高 AI 的采用率。可解释的 AI 还降低了与安全和公平的监管与声誉责任相关的风险。

越来越多的这类解决方案不仅能向数据科学家显示模型的输入和输出 , 还能解释为什么系统选择特定的模型 , 以及增强型数据科学和 ML 所采用的技术。

在训练 AI 模型时 , 偏见一直是一种长期存在的风险。它们可能是基于种族、性别、年龄或地点的偏见。此外 , 还有时间偏见 , 对特定数据结构的偏见 , 甚至在选择要解决的问题时的偏见。可解释的 AI 解决方案已经开始识别这些和其他潜在的偏见来源。

可解释的 AI 技术也可以识别隐私侵犯风险 , 并且它可以选择使用隐私感知机器学习 (PAML)、多方计算和各种同态加密来识别隐私侵犯风险。

### **该趋势如何影响您的组织和技能 ?**

多样化是可解释 AI 的关键基础 , 因为 :

- 数据多样化是客观看待问题并交付可信结果的必要条件。
- 算法多样化是权衡利弊的必要条件 , 特别是在权衡准确性与可解释性这个两难问题时更是如此。
- 人才多样化是组建成功的 AI 团队和 AI 道德委员会的必要条件 , 从而将声誉和业务风险降到最低 , 并确保 AI 安全性。

数据和分析负责人应该在培训和教育上进行投资 , 以培养相关技能 , 降低黑盒模型的风险。

其中应该包括：

- 如何设计可解释的数据科学和 ML 模型，以及如何从一系列模型（从最不透明到最透明）中选择正确的模型透明度（考虑透明度对准确性的影响）。
- 如何在需要的时候选择正确的模型准确性，并在“事后”验证和解释这些模型，例如模型通用或模型特定的可解释性；全局（跨整个模型）或局部（特定输出）的可解释性。
- 各种各样的方法，比如生成可解释性，结合利用简单但可解释的模型与复杂但不可解释的模型。
- 探索最新的可解释性技术，比如 DARPA 跟踪的技术，或者商业供应商提供的技术。
- 可视化方法，用于在训练和解释机器学习算法时查看和理解数据（例如，将数据中的关联或异常值可视化）。
- 用于理解和验证最复杂预测模型类型的技术，如敏感性分析、代理模型和“去除某一自变量的分析法”（LOCO）。
- 数据科学家的沟通和同理心技能，以检测用户对可解释性和成功采用 AI 的态度与需求。
- 建立 AI 道德委员会和其他负责 AI 安全性、公平性和道德的团体。这些委员会应该包括以声誉和诚信著称的内部和外部人士。

## 用例

企业以各种方式向分析和 BI 平台以及数据科学和 ML 平台提供可解释的 AI。

商用产品包括：

- 数据科学平台，如 DataRobot Labs 和 H2O.ai，它们可以用自然语言自动生成模型解释。
- [tazi.ai](https://tazi.ai) 为非技术业务用户提供模式可视化内容（例如损/益预测模式），帮助他们以交互方式通过解释调查可能的模式。

- DarwinAI 提供 Generative Synthesis 技术，该技术为深入洞悉神经网络性能提供了一个工具。
- ZestFinance 专注于金融行业，提供准确、透明的贷款 AutoML 模型。
- Salesforce Einstein Discovery 解释了模型的发现，并提醒用户注意数据中的潜在偏见。

除了提高模型可解释性的供应商技术之外，市场上还出现了许多解决 AI 可解释性挑战的计划。

- 美国国防高级研究计划局 (DARPA) 已经确定，AI 系统目前无法向人类用户解释其决策和行动，这限制了 AI 系统的有效性。DARPA 通过运行可解释的 AI 程序来创建一套 ML 技术，从而在提高可解释性方面发挥关键作用：
  - 生成更多可解释的模型，同时保持高水平的学习性能（预测准确性）。
  - 使人类用户能够理解、适当地信任和有效地管理新一代 AI 伙伴。

其他计划包括：

- 加州大学伯克利分校、加州大学洛杉矶分校、麻省理工学院、俄勒冈州立大学、罗格斯大学、SRI 研究所和 PARC 等机构正在引入可解释性解决方案，以检验深度神经网络 (DNN) 的某些方面，以解释 DNN 表示，并试验学习模型。<sup>1</sup>
- LIME (Locally Interpretable Model-Agnostic Explanations) 是一种开源技术，它分析数据样本的输入，并观察预测结果是如何变化的。<sup>2</sup>
- SHAP (SHapley Additive exPlanations)<sup>3</sup> 是一种开源技术，它统一了本地可解释性的六种不同的方法。<sup>4</sup>
- arXiv 上发布了很多有关可解释性的研究报告。<sup>5</sup>

## 建议

数据和分析负责人应该分配职责，聚焦 AI 可解释性的以下方面：

- 在评估分析、商业智能、数据科学和机器学习平台时，评估 AI 的可解释性功能。
- 使用选定供应商提供的功能来建立信任，并更广泛地采用自动生成的模型和洞察力。

- 针对每个用例，权衡准确性与可解释性。
- 开发决定何时需要可解释性的治理方法，以及评估权衡的指导方针。
- 知道并非所有 AI 解决方案都需要可解释性，但那些影响用户采用、监管和风险的解决方案需要可解释性。
- 建立问责制，以确定和实施每个用例的数据、算法和输出的信任级别与透明度。

### 相关报告推荐

“Build Trust With Business Users by Moving Toward Explainable AI”

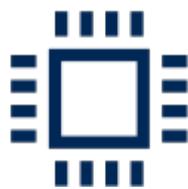
“Build AI-Specific Governance on Three Cornerstones:Trust, Transparency and Diversity”

“Seek Diversity of People, Data and Algorithms to Keep AI Honest”

“Predicts 2019: Digital Ethics, Policy and Governance Are Key to Success With Artificial Intelligence”

## 趋势 8：数据和分析领域的区块链

新兴



到 2021 年，大多数获许可的区块链用例将会被账本 DBMS 产品所取代。

区块链潜力无穷。它提供支持加密的数据不变性，并在参与者网络中共享。随之而来的是极高的复杂性。您的内部业务流程能够在您的网络中有效共享。而基于区块链的系统不是您的记录系统，这意味着您需要在数据、应用和业务流程方面开展大量的集成工作。最后，该技术还没有成熟到能够在现实世界中实现生产级可扩展性。

**为什么立即行动？**

区块链和分布式账本技术的核心价值主张是在一个基本上不受信任的参与者组成的网络中提供分散的信任。这种信任的基础是所有交易或状态变更都是不可变、可验证和公开记录的。这意味着，区块链网络中不存在信息不对称，因为所有的网络参与者都拥有完美的信息。对分析用例的潜在探索意义重大，特别是那些依赖参与者关系和交互的分析用例。

### **该趋势能提供哪些支持？**

区块链技术解决了数据和分析领域的两个挑战。首先，区块链提供了完整的资产和交易沿革。其次，区块链为复杂的参与者网络提供了透明度。

共享账本通过交互、互联的节点图的形式公开整个参与者网络。

这样，您就能识别：

- 哪些参与者相互交互。
- 第一次交互如何潜在地触发第二次或第三次交互。
- 交互如何随时间变化，交互是持久的还是短暂的。
- 谁是最重要的（也可能是最有影响力的）网络参与者。

您可以通过分析节点图，来了解通过网络流动的货物和数据及其对条件变化的敏感性（参见“Making Big Data Normal With Graph Analysis for the Masses”）。在供应链环境中，优化建模和模拟将变得更加丰富。虽然这些方法对于分析来说并不新鲜，但是基于区块链的网络大大简化了所有参与者的数据访问和一致性。

审计、产品沿革和欺诈分析等领域已经完成了将图形分析功能应用于分布式可信区块链网络的早期工作。数据共享和协作是一种未经测试但很有前途的潜在技术用途。

数据和分析领域有许多阻碍企业采用区块链的因素，包括：

- **数据管理能力不足**：区块链不是独立的数据存储。由于区块大小的限制（一个区块可以保存多少数据）和其他类型的分布式账本的扩展限制，大型数据集必须链下存储在别的数据存储中。表示数据的区块同时持有数据资产的校验和与统一资源标识符（URI）。校验和保证数据没有被篡改，而 URI 是指向实际数据集的指针。

- **数据生命周期管理尚不完善**：指向数据的指针可以被撤销或更新，链下文件也可以更新，但是实际的数据资产在下载或声明后不能被撤销。换句话说，没有与区块链实施集成的数字版权管理。目前至少有一家公司 (BigchainDB) 致力于解决该挑战。
- **对比特币之外的分析仅提供有限支持**：虽然区块链是一种很有前途的辅助数据和解析的技术，但这项技术仍处于起步阶段。目前基于区块链的分析工作主要专注于检测比特币区块链中的欺诈行为，主要是洗钱行为。
- **技术成熟度和可互操作性**：四到五种主要的区块链技术成为主流还需要几年的时间。在此之前，技术最终用户必须与主要客户或网络指定的区块链技术和标准相集成。其中包括与现有数据和分析基础架构相集成。这些集成成本可能会超过区块链带来的任何潜在收益。

### 该趋势如何影响您的组织和技能？

采用区块链的组织需要在分布式计算环境中重新构建它们现有的集中式业务流程。围绕数据、应用和流程集成的传统技能对于区块链的成功发展至关重要。企业还需要培养区块链特定的应用开发技能，包括智能合同。竞争性区块链实施项目之间缺乏可互操作性使得这项工作变得更加复杂。在区块链技术架构变得更稳定之前，组织应该需要支持多个区块链实施项目。

### 用例

**审计和产品沿袭**：所有供应链参与者共享的不可撤销的数据集也扩展了整个网络的可视性。这增加了可监控的表面积，提高了产品来源的可追溯性，加深了对产品来源的了解。Everledger 和 Provenance 等企业引领了这一领域的早期研究，前者追踪钻石和其他昂贵资产的全球流动，后者瞄准从印尼到日本的鱼类供应。

**欺诈分析**：公共分布式账本可以通过向所有各方提供相同的数据集，增强欺诈分析和风险管理。一方可以迅速标记检测到的欺诈活动并将其分发到整个网络，在其他参与者成为受害者之前通知他们。任何涉及多方交易或协调的行业（如医疗或保险行业）都可能受益于区块链或分布式账本提供的欺诈分析可能性。

**数据共享与协作**：将外部数据引入内部流程时，您必须将外部数据视为完全未经验证且不可

信的数据。当数据来源于外部治理模型时，这是一种合理的（尽管有些偏执）方法。数据的内容和情境不在您的控制范围内，它们会影响描述、共享和使用数据的方式（参见“Apply Gartner’s Information Capabilities Framework to Achieve Algorithmic Business Success”）。

在共享或自描述数据模型的支持下，区块链实施项目保证了不可撤销性和一致性，允许您将外部数据视为内部数据。您可以跟踪数据创建和修改的来源，从而保证数据的沿袭和来源。

## 建议

- 通过重点关注数据管理基础架构和区块链技术之间的功能错配，将区块链技术定位为辅助现有数据管理基础架构的技术。
- 通过组建一支试点团队，利用区块链在不变性、容错和交易透明度方面的优势，与业务负责人一起探索独特的用例。
- 灵活性设计。考虑到区块链技术早期的不稳定状态，在可能的情况下，围绕供应商特定的平台引入抽象层，借此限制您的长期承诺。

## 相关报告推荐

“Building Blockchain Into Your Data and Analytics Program”

“Debunking the Top 3 Blockchain Myths for Data Management”

“Amazon QLDB Challenges Permissioned Blockchains”

## 趋势 9：持续智能

智能



可扩展



到 2022 年，超过一半的关键新业务系统将会采用使用实时情境数据来改善决策的持续智能。

持续智能结合利用了数据和分析与事务性业务流程和其他实时交互。它利用了增强分析、事件流处理、优化、业务规则管理和 ML。

### 为什么立即行动？

长期以来，组织一直在寻求实时智能，目前已经有一些系统能为有限的一些任务提供实时智能。然而，随着以下技术的进步，在更大的范围内实施这些系统（Gartner 称之为持续智能）最终是可行的：

- 越来越多可用的增强分析，包括 ML、AI 和流分析；决策管理软件；以及时间序列 DBMS 软件。
- 低成本、高性能的 CPU、GPU、内存、存储、网络、云计算和移动设备。
- 来自无所不在的物联网 (IoT) 的廉价的传感器数据。

### 该趋势能提供哪些支持？

持续智能包含以下三个要素：

1. **情境感知**基于对流数据的连续（始终在线的）摄入，保持实时或近实时状态。
2. **主动行为**指的是当系统检测到需要注意的情况时，它可以推送警报、更新仪表盘或触发自动响应（而不只是被动地等待员工或应用查询或提取信息）。
3. **规定行为**指的是系统告诉您该做什么。它为人工决策或自动化流程的决策自动化提供决策支持（换句话说，它不只是通知您发生了什么，还会让您作出适当的响应）。

持续智能提高了各种运营决策的质量和精度，因为它将更多种类的可信数据融合到了计算决策所用的算法中。它与实时和近乎实时的决策相关。在制定这些决策时，了解当前的情况（或过去几秒或几分钟内发生的事件）是有好处的。

系统能够快速处理大量数据，避免人们超负荷工作。相比人工，它们能够运用规则和优化逻辑评估更多选项。

### 该趋势如何影响您的组织和技能？

持续智能对数据和分析团队的影响比这里列出的其他趋势的影响更大，因为它直接影响事务处理，面向的是客户、物流、B2B 和其他运营系统。数据和分析团队必须与应用架构师、应用开发人员、业务流程管理分析人员和业务分析人员合作，以设计、构建、部署和维护系统。

## 用例

持续智能的初级独立版本已经无处不在。您的移动设备上的导航系统是一个持续智能系统，它为您选择路线提供实时的建议。

一般来说，市场上有两种类型的持续智能系统：

- **主动推送系统**：人们被信息所包围，因此，他们需要实施异常管理。始终在线的持续智能（监视）系统全天候运行，监听发生的事件。当它们检测到需要响应的威胁或机会情境时，它们会更新仪表盘、发送警报或触发自动化响应。这些都是主动的推送，因为系统启动了响应。它们可能启动一个进程或向机器发送一个控制信号。实时仪表盘提供情境感知。
- **按需型系统**：根据定义，所有持续智能系统在某些流程中都是主动的。然而，在一些其他的流程中，几乎所有系统都是被动的（按需型系统）。当业务流程达到决策点，或者当有人选择触发某个流程或决策（例如贷款批准或下一个最佳产品）时，系统会调用按需型实时分析。

通过使用持续智能，企业获得的主要业务收益包括：

- 将全方位的客户视图转换为实时的全方位视图，以提供更精确、更有效的客户报价和更出色的客户支持。
- 实施基于条件的预测性维护，以延长机器的寿命，同时避免过早更换零件，减少浪费。
- 为航空公司、铁路公司、货运公司和航运公司提供企业神经系统，以监控和优化资源调度决策，同时提高客户满意度。

## 建议

数据和分析负责人应该：

- 与应用架构师、应用开发人员、业务流程管理分析人员和业务分析人员协作，共同制定集中式计划并设计、构建、实施和维护系统。

## 相关报告推荐

“Innovation Insight for Continuous Intelligence”

“Building Your Continuous Intelligence Capability for Digital Transformation”

“Make Your Customer Engagement Hub Real Time With Continuous Intelligence”

## 趋势 10：持久内存服务器

可扩展



到 2021 年，永久性内存将占内存中计算领域内存 GB 消耗量的 10% 以上。

### 为什么立即行动？

现在，大多数数据库管理系统都使用内存数据库结构，包括基于行和/或基于列的数据库结构。在大多数情况下，内存大小是有限的，您不可能把所有数据都保存在内存中。要么由 DBMS 软件决定在内存中保存什么，要么由用户定义。有些服务器确实有很大的内存（最高 64TB），但是，这些服务器需要多个处理器，因此非常昂贵。

在延期很长一段时间后，基于 3D Xpoint 内存技术的 Intel Optane DC Persistent Memory 将于 2019 年年中作为可直接寻址的内存全面上市。虽然它无法取代 DRAM，但它是一个 NVDIMM 中完全可寻址的大容量内存（目前最高达到 512GB）。与所有新的内存技术一样，硬件的价格需要几年的时间才能实现规模化，而且还需要经过几次迭代才能进一步降低成本。DBMS 供应商也需要几年的时间来修改他们的软件才能利用该技术。

许多 DBMS 供应商正在试验持久内存，一些产品（包括 Aerospike 和 SAP HANA）已

经做了必要的修改。尽管该软件现在已经可用，但它只能在选定硬件供应商的测试系统和 Google Cloud Platform 上使用。

### **该趋势能提供哪些支持？**

IT 所有领域的现代系统可用和需要的数据量正在迅速增长，将数据实时转换为价值的紧迫性也在以同样的速度增长。新的服务器工作负载不仅要求更快的 CPU 性能，还要求更大的内存和更快的存储。过去，DRAM 是一种昂贵但可靠且可按字节寻址的内存解决方案，但它仍然缺乏更便宜、更密集（但更慢）的非易失性 NAND 闪存的那种经济效益，无法充当可按区块寻址的存储。这种新的持久 NVDIMM 并不会取代 DRAM，而是可与 DRAM 搭配使用，优化软件环境。

Intel DV Optane 持久内存有两种运营模式：

1. **内存模式**：作为一个更大的内存池，即插即用。
2. **应用直接模式**：持久内存池，需要修改和优化应用。

在内存模式下，您不需要对软件进行任何变更，就可以使用更大的内存。但是，内存模式不包括持久内存。尽管使用更方便（因为不需要修改），但是它失去了持久内存的高可用性（HA）优势，因为它依然是不稳定的。

在应用直接模式下，您也可以使用大内存，并且是持久内存，发挥持久内存的真正优势。此模式要求供应商修改其软件以利用持久内存。这将对 HA、灾难恢复以及系统在故障后如何重启产生深远的影响。

### **该趋势如何影响您的组织和技能？**

使用持久内存时，您不需要或几乎不需要变更您的技能，因为必要的变更都发生在 DBMS 软件中。随着时间的推移，成本将减少，持久内存服务器将变得广泛可用，内存 DBMS 的采用率也将随之增长。对组织的影响体现在以下方面：整合、提高内存系统的 HA 效率、提高性能，以及支持新的架构，如混合事务/分析处理（HTAP）。

### **用例**

持久内存有很多不同的用例，包括：

- **虚拟化**：大多数软件虚拟化（如 VMware）都需要内存。持久内存支持更大型、更高效的虚拟化环境。这不仅适用于服务器虚拟化，也适用于桌面虚拟化系统。
- **DBMS 和数据网格**：数据网格已经是内存架构，越来越多的 DBMS 开始使用内存技术。持久内存允许大部分或所有数据保存在内存中，这不仅可以提高性能，还可以简化 HA，同时降低 HA 系统的重启影响。
- **分析**：现在很多分析供应商利用内存结构来提高性能。借助持久内存，这些系统能够增加内存中的数据量，从而进一步提高性能。使用 ML 算法的增强分析尤其如此，因为它需要大量数据。

## 建议

对数据管理解决方案感兴趣的数据和分析负责人应该：

- 请您的软件供应商确定他们在利用持久内存方面采取的行动，以及持久内容的可用时间。
- 评估基础架构中持久内存服务器的使用情况（当它们可用时），以了解其对于 HA 架构简化、虚拟化应用增长、服务器整合以及潜在成本节省方面的影响。

## 相关报告推荐

“Determining the Data Center Opportunity Created for 3D XPoint Persistent Memory”

## 对成功至关重要的非技术趋势

数据和分析不再处于业务的外围。它们是组织服务客户和优化业务流程的核心，也是新的颠覆性业务模式和收入流的基础。然而，为了发挥广泛的业务影响力，数据和分析负责人必须将这些投资扩展到单个部门和项目之外，以增强企业内外每个人的能力。

培养专注且由业务驱动的领导力和、营造由数据驱动的文化，以及找到正确的组织模型，这些措施都将发挥关键作用。这意味着您要比以往任何时候都更紧密地与企业内外的团队合作。扩大数据和分析团队的规模和覆盖范围，为这些团队带来更多适合的技能，并以一种跨业务的分布式方式扮演更多角色是成功的必要条件。

您必须进行特定的投资，以创建新的角色和职责（例如首席数据官），将数据和分析投资与战略性业务结果和价值联系起来。培养数据科学、机器学习、人工智能和数据工程领域的新技能，以及组织内每个人的数据素养和熟练度，是您在数字时代的新文化要求，也是组织成功的推动因素。您有必要建立新的工作方式和新的数据驱动型方法，以利用复杂多样的数据资产；同时，您还需要建立各种多样化的思维模式，激发创造力和创新力。

重要的是，为了让分析变得无处不在，您必须将数据和分析计划最初的成功经验应用到业务的所有方面。您必须根据业务线优先事项和结果，投资数据和分析作为战略优先事项，并将其推广到企业的各个角落。这决定了您能否实现大多数数字业务目标。

数据和分析负责人应该像关注技术趋势一样关注和投资非技术趋势。

图 2 列出了对于数据和分析计划的成功至关重要的非技术趋势。

图 2：对于数据和分析计划的成功至关重要的非技术趋势



#### 推荐阅读：

“The Future of Data and Analytics: Tales and Trends From the Center to the Edge”

“Leadership Vision for 2019:Data and Analytics Leader.”

“Getting Started With Data Literacy and Information as a Second Language: A

Gartner Trend Insight Report”

“Fostering Data Literacy and Information as a Second Language: A Gartner Trend Insight Report”

“Toolkit: Enabling Data Literacy and Information as a Second Language”

“Survey Analysis: Third Gartner CDO Survey — How Chief Data Officers Are Driving Business Impact”

“How Data and Analytics Leaders Can Create Effective Influence and Communication Strategies”

## 证据

- 1 [“Explainable Artificial Intelligence \(XAI\),”](#) DARPA (pdf)
- 2 [“LIME,”](#) GitHub
- 3 [“SHAP,”](#) GitHub
- 4 [“A Unified Approach to Interpreting Model Predictions,”](#) NIPS Proceedings
- 5 [“Explaining Explanations to Society,”](#) [“Interaction Design for Explainable AI,”](#) [“Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions”](#)

© 2020 Gartner, Inc. 和/或其附属公司版权所有。保留所有权利。Gartner 是 Gartner, Inc. 及其关联公司的注册商标。未经 Gartner 提前书面许可，不得以任何形式对本出版物进行复制或分发。本出版物中包含 Gartner 研究机构的观点，不应被理解为事实陈述。本出版物中所含信息均来自可靠来源，但是 Gartner 并不保证这些信息的准确性、完整性和充分性。尽管 Gartner 研究中可能包含相关法律和财务问题的讨论，但 Gartner 不提供法律或投资建议，而且其研究不应被理解为或用作法律或投资建议。您对本出版物的访问和使用受 [Gartner 使用政策](#) 的约束。Gartner 的研究在独立性和客观性上享有很高的声誉。Gartner 研究由其研究机构单独进行，未接受任何第三方的输入，亦不受任何第三方的影响。有关更多信息，请参阅 [“独立性与客观性指导原则”](#)。

# Gartner

© 2018 Gartner, Inc. 和/或其附属公司版权所有。保留所有权利。

