数据可观察性的权威指南

——用于数据分析和人工智能

The Definitive Guide to Data Observability for Analytics and AI

译者：Rain

目录

引言

爆炸性的数据供需正在将现代数据管道推向崩溃点。企业数据消费者希望使用来自更广泛来源的更多数据，通常是实时的，以改进决策和优化操作。但是数据团队很难构建、构建和操作数据系统，以满足这些快速扩展的业务需求。

新的工具和平台，加上在工程和运营方面的更大投资，只能部分缓解痛苦。现实情况是，大多数企业数据团队仍然会花费大部分时间来解决日常操作问题。随着大量的**数据量、数据管道的复杂性**和新技术共同淹没数据团队的能力，**破坏数据系统的业务价值**，这个问题只会变得越来越严重。

数据可观察性的范式试图解决这个前所未有的数据复杂性的新世界。**数据可观察性**建立在其前身的**应用性能监测(APM)**的基础上，提供了一种系统的方法。它试图**跨应用程序、数据和基础设施层来监视和关联数据事件**。通过这样做，它使**企业所有者、DevOps 工程师、数据架构师、数据工程师和站点可靠性工程师**能够**检测、预测、预防和解决问题**——有时是以自动化的方式——否则将打破生产分析和人工智能。

为了成功地实现数据可观察性，数据分析部门的领导者必须收集**需求并对需求进行优先排序**，然后选择一个全面的数据可观察性产品，以最小化自定义集成工作。他们应该首先处理小型的、可实现的可观察性项目，招募一个跨职能的贡献者团队，以关注关键的痛点，如绩效和效率。早期项目的成功可以导致更雄心勃勃的可观察性工作——前提是业务和 IT 领导者继续替换和淘汰重复的旧工具。

数据管道的演变

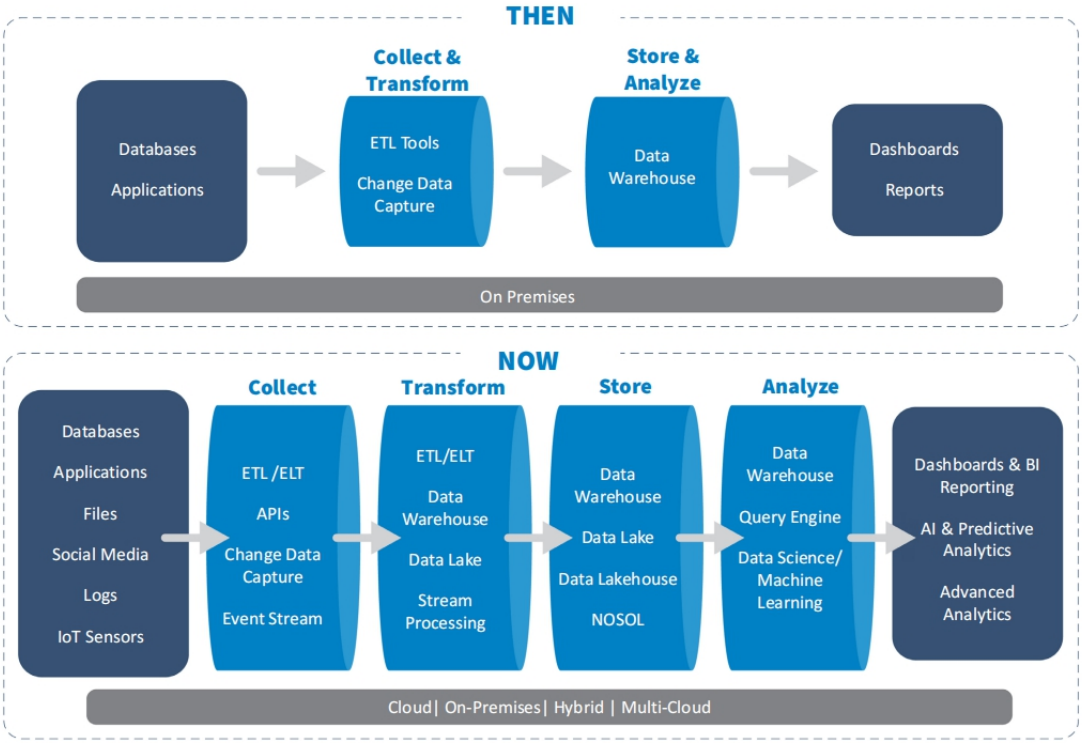
多年来，企业数据管道为业务分析提供了严格但相对稳定的需求。由商业智能(BI)分析师组成的小团队需要定期对其销售渠道、财务状况、库存水平和其他功能指标进行历史测量。他们依赖数据工程师来构建基本的数据管道，其中一些数据提取、转换和加载(ETL)和更改数据捕获(CDC)工具从数据库和应用程序中摄取结构化数据。数据工程师使用这些工具将数据转换并加载到批处理作业的数据仓库中，通常是一夜之间。分析人员使用传统的 BI 软件来生成仪表板和报告。

日益复杂：大约 15 年前，数据的供应和需求激增，此后，数据管道逐渐变

得更加复杂。越来越多的企业数据消费者，从运营管理人员到数据分析人员到数据科学家，开始使用新算法从新的数据源中生成新的情报和分析，通常是实时的。这促使数据团队采用了新的工具和平台，从 BI 的 Snowflake 到 Databricks 和数据科学的亚马逊卫星制造商。这些工具和平台利用了云本地存储和计算基础设施，提供了前所未有的弹性和可伸缩性。与此同时，数据量一直在上升。

架构师和数据工程师现在使用大量的工具来构建数据管道：提取、加载和转换(ELT)工具、CDC、应用程序编程接口(api)和事件流系统，如 ApacheKafka。它们从社交媒体、IT 日志和物联网（物联网）传感器中获取结构化、半结构化和非结构化的数据。他们将这些数据转换并存储在数据仓库、数据湖、NoSQL，甚至流媒体平台中。云对象存储和计算引擎必须与遗留的内部部署系统集成，从而增加了复杂性。在这种混合和多云的混乱中，公司必须建立脆弱的数据管道，将数据输入到成倍增加的目标集，包括 BI 工具、人工智能(AI)和嵌入式分析工作流。

图 1 说明了数据供需爆炸对企业数据管道的影响



这种复杂性和不断上升的数据浪潮可能会压倒管理作为现代分析管道基础的基础设施、应用程序和网络的企业团队。它们与缓慢、笨拙的 Hadoop 数据湖斗争，由于数据重力，这些湖在内部持续存在。他们难以控制云数据平台的性能和成本，同时跨混合、多云环境维护集成视图。交付具有足够性能、可用性和可靠性的数据和分析管道变得越来越困难。简而言之，**数据管道的复杂性破坏了它**

们为业务提供的价值。

数据管道的复杂性破坏了它们为业务提供的价值。

这就是可观察性技术的切入点。它使数据团队能够监控现代分析和人工智能应用程序基础上的数据基础设施，并检测可能限制性能或导致中断的问题。有效的数据可观察性工具必须满足堆栈的每一层的需求。

- 基础设施层：平台工程师、DevOps 工程师和现场可靠性工程师(SREs)需要监控存储和计算可用性、利用率、性能及其对数据流的影响。如果没有对这些系统有足够的可见性，以及将活动与数据和分析管道问题关联起来的能力，工程师就会遇到操作问题、性能瓶颈和系统中断。
- 数据层：数据架构师和数据工程师需要监控数据库和网络应用程序，如 ApacheSpark 和 ApacheKafka，它们为现代数据和分析管道提供动力。为了提高处理吞吐量和最小化网络延迟，他们需要快速发现和解决问题的工具。否则，数据的及时性和质量不符合服务水平协议(SLAs)。
- 应用程序层：BI 分析师、数据科学家和业务经理需要理解性能问题的根源，这使得构建和使用分析工具和数据具有挑战性。传统的应用程序性能监控(APM)工具可以识别应用程序问题，但它们不能回答与数据相关的问题，比如为什么 ETL 作业会挂起。

在没有可观察工具的情况下，这些专业人士玩鼹鼠游戏，一个接一个问题，却让下一个问题几乎瞬间出现。这是一场持续的斗争。数据团队通过自定义各个工具，将监视视图拼接在一起。他们应用快速的新技术，如 DremioSQL 查询引擎，亚马逊运动流处理器，或 Apache Druid 分布式数据存储。但最终，他们仍然必须手动查找性能问题，并确定对系统可靠性以及在分析管道结束时的数据消费者和业务用户产生不利影响的中断的根本原因。

定义的数据可观察性

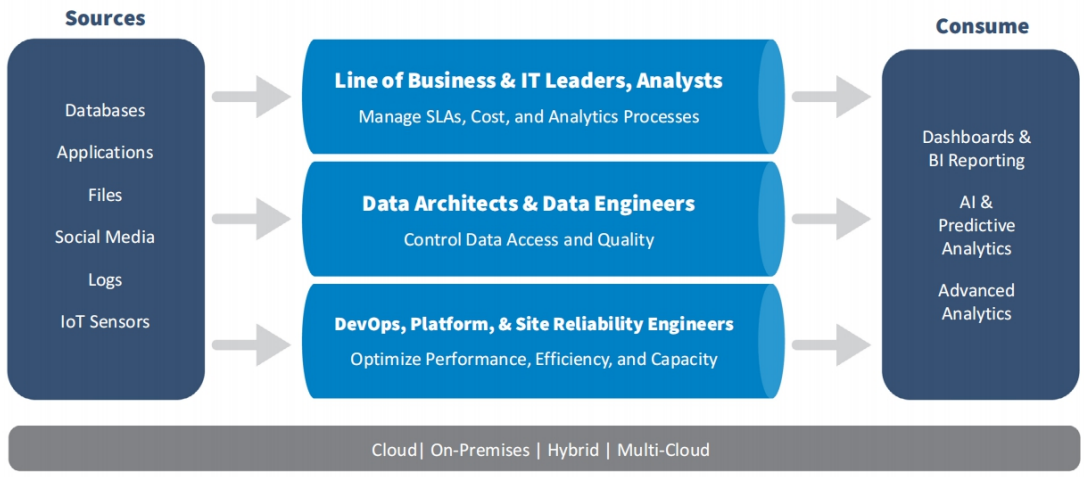
数据可观测性提出了一个系统的解决方案。它寻求改进对处理人工智能和分析工作负荷的所有元素的控制。可观察性使业务和 IT 组织能够监控、检测、预测、预防和解决整个企业中从源到消费的问题。**可观察性使用自动化和人工智能将跨多个服务器、节点、集群、容器和应用程序的数千个警报关联起来。它精确定位问题并帮助解决问题，因此系统可以扩大规模并开始再次满足期望**。可观察性还包括**数据质量控制**，以帮助减轻不准确的人工智能和分析输出不断上升的风险。这不是灵丹妙药，但这是一个很好的开始。

数据可观察性意味着业务和 IT 可以监控、检测、预测、预防和解决跨企业数据管道的问题。

具有数据可观测性：

- DevOps、平台和现场可靠性工程师可确保基础设施的性能、效率和容量。
- 数据架构师和数据工程师提高了数据访问性和质量。
- 数据团队可以满足或超过 SLAss。
- 业务领导、IT 领导和分析师的系列将改进决策、分析流程规划和成本控制。

图 2 说明了关键利益相关者如何使用数据可观察性。



用例

数据可观察性支持数据团队，并支持通常跨越堆栈的多个角色和层的用例。关键利益相关者的效率将显著提高，因为常规任务可以实现自动化、简化流程和改进

决策。

数据可观察性通过自动化任务、简化流程和实现更好的决策来提高员工的生产效率。

DevOps，平台和现场可靠性工程师

数据可观察性工具可以帮助进行性能管理、基础设施精简和容量规划

基础设施性能管理。DevOps、平台和站点可靠性工程师可以为内存可用性、CPU/存储消耗和集群/节点状态等指标配置可观察性监视器。它们可以**定义警报阈值和通知**，所有这些都可按实体类型排序——用户、应用程序、CPU、磁盘、分布式文件系统如 Hadoop 文件系统(HDFS)、处理器如 Spark、如 YARN 的调度程序、容器系统如 Kubernetes 等。此级别的粒度有助于识别数据流拥塞、中断和失控的用户或应用程序。然后，工程师可以进行故障排除，并深入研究工作和组件，这有助于解决问题。例如，它们可以自动调整服务队列、任务和容器的容量级别。

基础设施精简。数据集通常有倾斜度，这意味着大多数输入/输出（I/O）集中于一小部分数据。可观察性揭示了数据集的倾斜度，以帮助降低存储成本。例如，您可以按大小过滤或排序文件，以识别最近未访问过的大文件。将这些文件归档到 "冷存储"，如 Amazon S3 Glacier，可以节省成本并释放容量以适应增长，并释放容量以适应增长。您还可以重新平衡此基础结构层中的现有数据集，以支持提高数据层和应用层的性能。

容量规划。 DevOps、平台和现场可靠性工程师使用数据可观察性来测量和预测满足业务 SLAss 所需的资源。它们监视数据工作负载，以查明受约束的资源，或识别备用 CPU，并将任务重新分配给它们。人工智能驱动的特性可以帮助他们根据可用的容量、必要的缓冲区和预期的工作负载增长来计算未来的容量需求。它们通过预测何时缺乏资源或失控的工作负载将创建性能红线来设置性能范围。

数据架构师和数据工程师

数据可观察性可帮助数据团队管理管道性能和数据质量，并随着时间的推移提高体系结构效率和有效性

数据管道性能管理。数据架构师和数据工程师必须自动收集数千个管道事件，并将它们关联起来，识别和检查异常或峰值，然后使用这些发现来预测、度量、预防、故障排除和修复各种问题。例如，它们必须密切跟踪和调整分布式调度器如何为本地数据集中的集群节点分配作业。他们需要监控 Apache Spark 或 Hive 服务器的读写 I/O，并将这些指标与内存或 CPU 利用率和执行时间相关联，作为另一个例子。数据可观察性工具提供了这样的视图来帮助和推荐调优性能的方法，例如通过增强 CPU 或内存。

数据质量。领先的可观察性工具从数据操作手册中获取一页内容，并自动检查数据传输的准确性、完整性和一致性。它们创建规则来比较源与目标表或增量更新，然后标记警报、检查、删除和协调的不匹配。违规行为可能包括空值、重复记录、更改的模式或不匹配的值范围。数据可观察性工具还可以端到端跟踪沿袭，并与源工具和目标 BI 工具进行集成。这种操作数据质量检查确保数据管道满足预期。但是，它们并没有消除对专用数据质量解决方案的需要，这可以解决特定行业的合规法规等问题。

架构设计。数据架构师和数据工程师还必须从日常火拼中抽身出来设计更好的架构。他们可以将新发现的见解应用于管道性能和利用趋势中，将他们今天拥有的点和他们明天需要的点之间联系起来。他们可以根据观察到的工作负载行为、场景建模和影响分析，自信选择、部署和配置新的平台，如高性能的 Apache Druid。虽然在一个快速变化的世界中没有灵丹妙药，但数据可观察性降低了架构规划的风险。

业务线和 IT 领导者，以及分析师

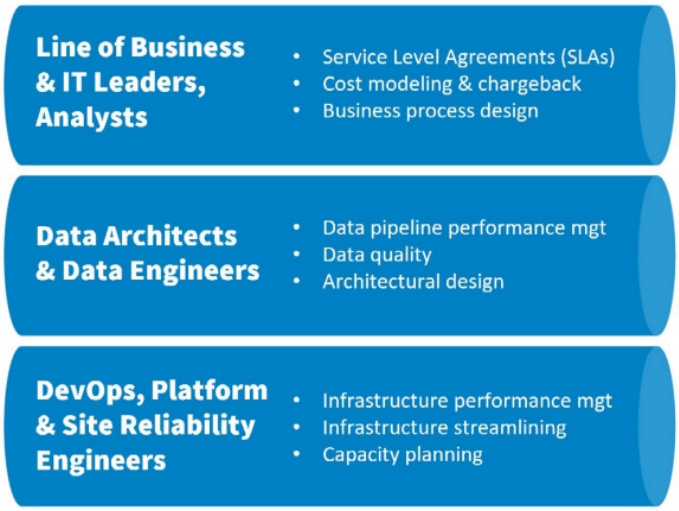
数据可观察性还使消费数据和监督数据使用的业务领导、IT 领导者和分析师的生活更轻松

服务级别协议(SLAss),大量的数据消费者继续要求能够实现它的数据架构师和数据工程师提供严格的延迟、吞吐量和正常运行时间服务承诺。虽然这种紧张局势不会很快缓解，但数据可观察性有助于双方进入更负责任的 SLAss。现场可靠性和平台工程师可以创建更准确的容量估计值。数据架构师和数据工程师设计得更快、更可靠的数据管道。业务领导、IT 领导和分析师可以更好地指导哪些 SLAss 是真正可行的。

成本决策和收费。业务线和 IT 领导者使用数据可观察性，以一种更准确、更细粒度的方式来管理数据分析的业务。他们可以与 BI 分析人员、数据架构师和数据工程师合作，基于聚合计算、存储、内存和数据传输需求等估计值，对其 SLAs 的隐含操作成本进行建模。他们可以按时间段、用户组、地理位置等来划分这些估计值，以帮助细化预算和退款决策。虽然没有单一的数据可观察性工具涵盖每一个可能的组成部分，但这些产品支持比以前可能的更明智的财务决策。

业务流程设计。出于各种原因，企业继续将分析和人工智能嵌入到其业务的更多方面，包括实现实时欺诈预防、客户建议或物联网预防性维护。它们需要增强而不是中断操作，这意味着它们需要对其数据工作负载 SLAs 有很高的信心。数据的可观察性很有帮助。业务领导、IT 领导、数据架构师和数据工程师可以协作设计创造性的分析-操作工作流，具有可接受的风险水平，理想情况下是合理的投资回报率。

图 3 按角色总结了这些用例。



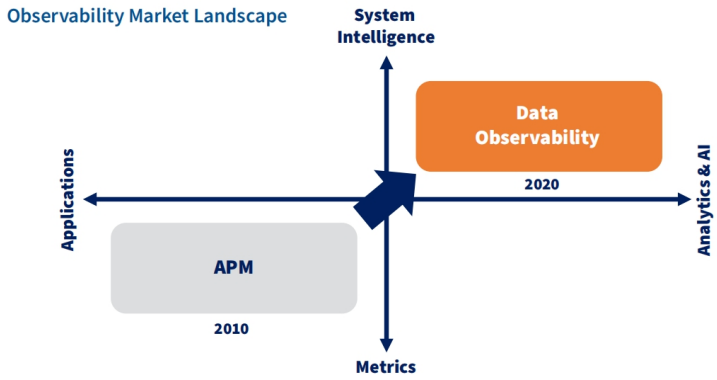
市场前景

数据可观察性扩展了企业多年来用于优化应用程序的 APM 工具的核心功能。传统的 APM 工具可以发现应用程序、监控它们的工作原理和诊断问题，通常需要借助人工智能和机器学习(ML)。它们跨微服务、服务器、容器和其他资源跟踪请求和关联事件。它们识别风险、减速、阻塞和故障，然后帮助修复它们，以改善和维护操作工作负载。

数据可观察性以两种方式建立在 APM 之上。首先，它将这些熟悉的 APM 函数应用于数据分析工作负载，而不是操作应用程序。其次，数据可观察性提供了跨数据、应用程序和基础结构层的更深层次相关性。这将将指标集成到系统智能中。例如，数据工程师可以使用可观察性来隔离数据层上的慢速 ETL 作业，然后与平台或站点可靠性工程师一起工作来隔离导致问题的存储配置错误。它们可以根据执行时间对 Spark 作业进行优先排序，然后深入到有问题的区域进行检查和调试。数据可观察性还可以帮助分析师和企业主控制其企业分析和人工智能堆栈的效率和成本。

数据可观察性将传统的 APM 功能应用于数据分析工作负载

图 4 比较了 APM 和数据的可观察性



数据可观察性与另外两个细分市场共享功能。首先是数据操作，它将 DevOps 和敏捷软件开发的原则应用于数据管道的创建和管理。数据可观察性解决了数据操作难题的监视、数据验证和沿袭部分，帮助确保数据质量和数据交付性能。其次，ITOps 工具（及其 AIOps 子段）提供、管理、监视和调优 IT 基础设施资源。数据可观察性解决了 ITOps 和 AIOps 的监测、诊断和补救方面。

提高数据可观察性工作

在开始实施可观察性计划之前，数据团队应该了解数据现代化作为采用驱动程序的作用。他们还应该考虑数据可观察性所面临的挑战及其预期的好处。

数据现代化

企业正在努力将其数据资产货币化，这推动了采用需要数据可观察性的大型共享平台。这有几个维度。许多面向业务的分析人员现在使用自动化和预打包的 ML 代码，这将 ML 用户基础扩展到了数据科学家之外。这些项目消耗大量的数据，包括社交媒体文本、物联网传感器，甚至卫星图像。它们提高了对速度和规模的需求，这反过来又促使数据团队通过基于弹性云基础设施的流媒体、数据湖和数据仓库平台来实现其架构的现代化。虽然云本地资源通常会提高性能和易用性，但它们与遗留的本地系统的链接又产生了需要注意的新风险和复杂性。

挑战

矛盾的是，复杂性也是可观察性计划面临的主要挑战。没有一个单一的商业解决方案可以覆盖所有温和环境的变化，这意味着数据团队必须定制他们的可观察性软件，或者接受比他们需要的更少的可见性。复杂性也困扰着监控任务：数据团队必须仔细配置和过滤他们的日志、追踪和指标，以提高信号与噪音的比率。如果前期没有充分关注这些细节，可观察性解决方案可能会造成混乱，损害生产力，并提高性能风险。数据可观察性的早期采用者，如 GE Digital，已经克服了

这些挑战，例如，将几个工具整合到一个更容易管理的单一监控平台上（见下面的案例研究）。

优势

当计划、实施和管理良好时，可观察性会带来几个好处。更低的延迟、更高的吞吐量和更准确的数据都有助于分析师和数据科学家从他们的分析和人工智能项目中获取新的价值。提高数据管道的可靠性、正常运行时间和问题解决方案可降低操作风险。监控工具的整合减少了管理开销，提高了数据团队的生产力，并创建了一个“人才桥”，从而减少了对培训或招聘的需求。数据体系结构还变得更加灵活、可伸缩和高效，并能够随着业务的需要而弯曲。

常见部署环境

数据可观测性为三种重叠类型的环境提供了最大的价值：内部部署的 Hadoop 数据湖、混合云和多云。每一个项目都带来了不同的挑战和需求的组合。

内部部署 Hadoop

数据可观察性可帮助数据团队管理在大型内部部署环境中持续存在的 Hadoop 数据湖。五到十年前，企业在 Hadoop 植入了一些分析数据和工作负载，即使云上出现了更易于管理、具有成本效益的替代方案，许多企业也没有完全放弃投资。因此，一些数据团队仍然在 Hadoop 上运行分析，需要帮助维护无数 Apache 开源组件的性能水平。与 MapReduce 相比，Spark 加速了 Hadoop 上的批处理，但通常需要仔细的监视、故障排除和调试，以满足生产分析延迟和吞吐量要求。数据的可观察性可以用来提高性能、可靠性和可伸缩性。

混合云

数据仓库和数据湖正在汇聚在云中的一组公共函数上，将高性能的 SQL 查询结构与高效、有弹性的对象存储进行配对。企业采用这些云数据平台，如 Azure 突触、数据库板和雪花，以减少管理麻烦和合并数据工作负载。但是数据团队仍然需要可观察性来密切关注云平台的性能，例如，以满足 BI 查询延迟的需求。它们还需要减少计算成本超支的风险，并维护与遗留的本地系统的所有必要链

接。

多云

随着企业数据团队获得在云计算方面的经验，他们寻求通过购买新的人工智能工具、云数据平台等来优化工作负载并满足专业需求。因此，除了遗留的本地系统之外，许多企业环境现在还包括两个甚至三个云服务提供商。可观察性可以帮助他们监督这些分布式拓扑，并维护高效、有效的数据管道。

案例分析

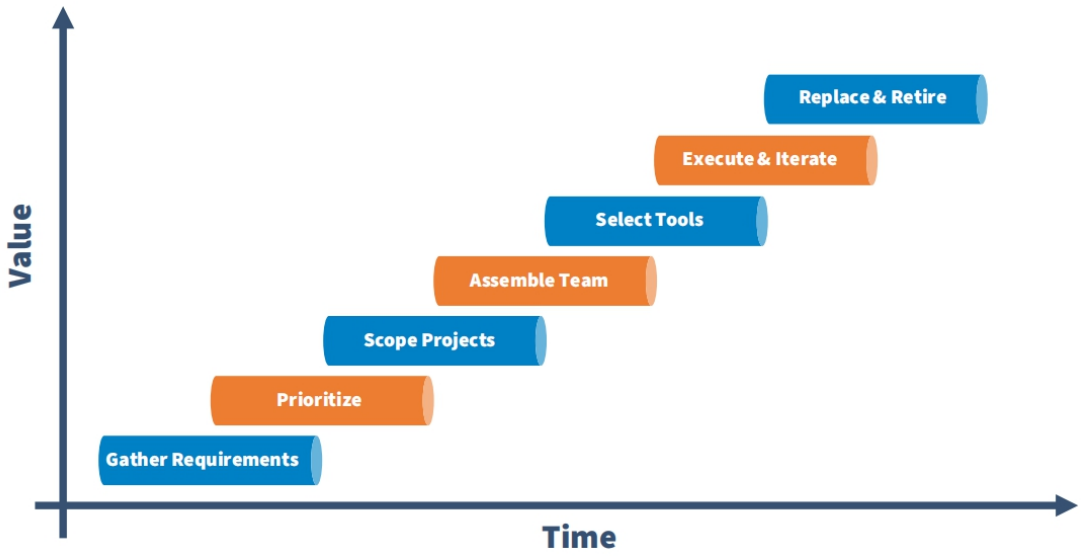
GE digital。GE 数字运行着全球第五大 Oracle 企业资源规划(ERP)和第三大 SAPERP 实现，利用了 1000 个 AmazonWeb 服务(AWS，AWS)核心。在接受可观察性之前，通用电气的数据团队无法理解和控制无数数据管道组件之间的交互作用，这些组件每分钟为操作和分析工作负载处理 600 万笔金融事务。这削弱了通用电气实现业绩目标的能力。“握手太多了，”首席数据官迪瓦卡尔·戈尔说，“一个错误的代价非常高。”

通用电气部署了一个来自 acceldata 的可观察性解决方案，以重新获得对其数据管道的控制。他们的数据团队集中了一些工具的监控活动，包括用于管理 Hadoop 和用于日志分析的开源 ApacheAmbari，到 Accel 数据上。他们驯服了长期存在的 ApacheSpark 处理性能问题，使工程团队将分析工作负载从有问题的 Hadoop 数据湖迁移到 AWS 上的单存储(以前称为 MemSQL)。Goel 估计，通过金融数据湖的努力，通用电气的运营成本每年减少了 3,000 万美元。它们的关键任务系统必须具有可观察性。

在接受可观察性之前，通用电气的数据团队无法理解和控制无数数据管道组件之间的交互作用，这些组件每分钟为操作和分析工作负载处理 600 万笔金融事务。

开始进行数据观察

数据可观察性之所以重要，是因为它解决了一个无可争辩的技术问题——即复杂性，这个问题只会在未来几年的严重程度上增加。企业必须采用一个良好的可观察性程序来降低而不是增加复杂性。要实现这一目标，需要一种有系统的、一步一步的方法，每一步都会为您的组织提供新的价值。图 5 显示了一个数据可观测性程序的阶段和相对值。



让我们考虑如何通过这些步骤取得快速、然后可持续的进展。

收集要求。数据可观测性的价值取决于它的整体方法。构建一个全面的需求清单，以涵盖跨基础设施、数据和应用程序层的用户、用例、技术组件和管道的广度。例如，您可能需要解决影响“实时”客户提供的生成和转换的 Spark 瓶颈。您可能会努力预测会导致云计算成本失控的数据科学工作负载飙升。在另一个级别上，您可能需要根据实际的基础设施资源消耗来创建或改进退款估计数。尽可能多地记录这些需求，包括潜在的未来需求。

优先排序。找出最痛苦和“可解决”的要求。您可以根据业务指标来评估疼痛——即由于事务处理性能、客户等待时间等而造成的收入或成本影响。“可修复”需求是那些最容易修复、风险最低、业务效益最高的需求。虽然需求差异很大，但较高的优先级需求往往来自于内部部署数据湖实现、混合云数据迁移和高级分析计划。

范围项目。开始搜索您的第一个项目，以满足这些最高优先级的要求。定义

您需要更改的功能，以及您应该如何更改它们。这个项目需要哪些人员、流程和技术变化？您可能需要雇佣或重新培训员工、重新配置数据管道、更改应用程序或实现新的工具。定义所有相关的任务及其依赖关系和风险，然后绘制可实现的项目里程碑，以改进和度量进度。通过从一个小的、更容易实现的项目开始，你就增加了快速获胜的几率，这表明了对更大组织的吸引力。

组装您的团队。接下来，您需要确定并招募必要的跨职能团队成员来推动您的项目。您需要来自业务领导者和 IT 领导者的执行赞助，因为他们可以分配必要的资源以确保成功。考虑建立一个卓越中心，由 IT 中心培训跨业务部门的各个利益相关者，以使用共同的策略、实践和工具来管理数据的可观察性。为您的第一个项目仔细选择一个小团队，与经理一起工作，在分析人员、数据架构师、数据工程师以及 DevOps、平台和/或现场可靠性工程师的队伍中确定正确的技术和业务主题专家。总的来说，项目团队应该拥有您的重点需求和用例的所有必要知识，其中可能是机器学习、数据仓库、ETL、流媒体等。

选择工具。很有可能，您的需求阶段确定了监控能力中只有一个全面的数据可观察性产品才能填补的空白。您可以根据几个关键的标准来评估潜在的产品。它是否对应用程序、数据和基础设施层的当前和近期未来组件提供了足够的可见性？它需要什么级别的自定义定义？它需要什么程度的培训，以及该工具的预期好处是否证明了成本和增加时间是合理的？此工具可以替代您所在环境中的其他工具吗？

执行和迭代。一旦您选择并实现了可观察性工具，您就可以开始项目执行了。一定要根据面向业务的基准测试来衡量每个项目的成功程度——无论它是否解决了这些 Spark 瓶颈、预测数据科学工作负载、改进退款等。更快的客户推荐对收入的影响是什么？通过更好的工作量预测或收回费用，您有多少钱能降低总体成本？根据您的回答，您可以完善您现有的项目，并扩展更成功的未来项目。一旦你证明了成功，你就可以获得行政赞助和资金来解决更多的战略项目。

更换并退休。一个典型的 IT 错误是采用了一种花哨的、易于使用的技术，但要保持其笨拙的前身，从而使生活比以往任何时候都更加艰难。不要让这种情况在数据可观察性中发生。为您的新可观察性产品构建并坚持一个明确的阶段性计划，以取代其他 APM 或 ITOps 工具中的这些功能。否则，您的团队将比以往任何时候都更分心，效率更低，从而对性能、数据质量等产生预期的影响。通过无情地精简和逐步淘汰旧的流程，您可以从系统中挤出额外的成本和效率。

数据可观察性为企业提供了在数据堆栈和组织的各个级别的关注。数据分析部门的领导者应该谨慎地选择自己的选择，以缓解数据访问、性能、质量、效率和成本方面的痛苦和风险。通过这样做，他们可以解决现有的问题，同时显著增加了在未来的分析和人工智能项目中可实现的优势。