acceldata

通过数据质量、资源效率和支出预测提高您的snowflake投资回报率

acceldata

为什么数据质量如此重要

(>)数据是现代数据驱动型企业的命脉。

(>)高质量的数据——准确、完整、一致和及时的数据——是最重要的。

(>)然而，即使是最先进的企业也会受到数据的影响不准确、有重复或缺失的记录、不一致的结构/模式、不透明的沿袭等等。

(>)质量差的数据会造成一大堆问题。

坏数据输入，垃圾信息输出。您会在执行报告中得到错误的数字，在仪表板中得到不稳定的图表，或者分析数据操作出现故障。直到晚上 11 点，全大写的电子邮件或 Slack 消息到达，您才知道为什么关键任务工具脱轨，或者老板演示文稿中的季度数据没有加起来。

根据《哈佛商业评论》调查的 42% 的高管表示，糟糕的数据质量是组织产生可操作的业务洞察力的第一大障碍。

不断的灭火，失去了敏捷性。对于选择不正面解决数据质量问题的组织，该决定几乎总是适得其反。他们的数据团队最终被拉到各个方向来修复损坏的仪表板、应用程序和管道。

与优先考虑数据质量相比，他们最终总是花费更多的时间和金钱来解决问题。这也使他们的数据工程师工作过度，因不断收到警报而感到疲倦，并因无法专注于创造价值的项目而士气低落。

数据驱动文化的解体。当员工不信任数据时，他们拒绝重用现有的数据管道和应用程序，而是要求全新的。构建这些对于您的数据工程师来说非常耗时，并且会增加您的存储成本。它还导致数据孤岛和暗数据池的激增，加剧了数据质量问题。

比缺乏有效的数据重用更糟糕的是不受信任的数据对决策的影响。商业领袖开始忽视可靠的定量支持的洞察力，转而支持直觉和轶事，您的企业一直在精心培育的数据驱动文化开始分崩离析。

对底线不利。数据质量具有巨大的财务影响。根据德克萨斯大学的一项研究，将数据质量和可用性提高 10% 的财富 1000 强公司平均每年可额外获得 20 亿美元的收入。对于未能提高数据质量的公司来说，这是一个巨大的错失良机。

Gartner 分析师 Ted Friedman 表示：“随着组织加快数字 [转型] 努力，糟糕的数据质量是导致信息信任和商业价值危机的主要原因，对财务业绩产生负面影响。”

是什么导致数据质量问题?

(>)**数据质量可能因多种原因而下降.**

(>)**架构更改可能会破坏为分析应用程序和仪表板提供数据的流程。**

(>)**API 调用可能会失败，从而中断数据流.**

(>)**手动、一次性的数据检索可能会产生错误和隐藏的重复数据池。**

可以出于充分的理由复制数据，例如提高查询性能。但是，如果没有强大的数据治理，这最终会导致大量昂贵的数据孤岛令人困惑。

从本地基础架构迁移到云也可能带来一系列新的数据质量和管理挑战。缺乏对整个数据生命周期的统一视图也会造成不一致，从而降低数据质量。

最后，今天几乎所有企业都面临一个问题：数据规模。根据 IDG 的一项调查，企业收集和存储的数据量正以惊人的速度增长——每月增长
高达 63%。数据源的数量也很大：平均公司有 400 个，20% 的公司有 1,000 多个来源。

现代数据堆栈的每一层都存在着一股数据工具的浪潮。公司不乏选择，从事件和 CDC 流媒体平台、ETL/ELT 工具、将洞察力推送到业务应用程序的反向 ETL 工具、数据 API 和可视化工具、实时分析数据库等等。

这些数据工具中有许多是点解决方案，是市场的早期进入者。尽管每个都有其优点，但试图从这些未集成的工具中拼凑出一个堆栈有助于创建碎片化、不可靠和损坏的数据环境。

为什么传统数据质量策略会失败

多年来，公司一直试图解决数据质量问题，通常是通过手动创建**数据质量策略和规则**，通常由主数据管理 (MDM) 或数据治理软件管理和执行。

Informatica、Oracle、SAP、SAS 等 MDM 供应商已经存在了几十年。他们的解决方案早在云或大数据出现之前就已经诞生和成熟。

毫不奇怪，这些过时的软件和策略无法扩展以适应当今更大的数据量和不断变化的数据结构。脚本和规则必须由人类数据工程师一一创建和更新。当警报响起时，您的数据工程师还需要手动检查异常、调试数据错误和清理数据集。这既费时又费力。

传统数据质量方法失败的一个很好的例子是**手动 ETL 验证脚本**. 数据工程师长期以来一直使用它们来清理和验证最近摄取的数据。应用于静态数据，ETL 验证脚本易于创建且灵活，因为它们可以用大多数编程语言编写并支持任何技术、数据系统或流程。

但是，手动 ETL 验证脚本通常不太适合当今企业数据环境的容量、速度和动态特性。获取流数据。事件和消息流的容量可能太大、太动态（具有不断变化的模式）和太实时，以至于 ETL 验证脚本无法工作。这些脚本只能批量处理数据，每次更改数据结构时都必须手动编辑。

这会导致显着的验证延迟。而这种延迟对于正在进行数字化转型的公司来说是不可接受的，因为它排除了实时客户个性化、数据驱动的物流、欺诈检测和其他内部操作、实时用户排行榜等用例。

除了实时数据之外，手动 ETL 验证脚本还有其他问题。对数据架构、系统、模式或流程的任何更改都将迫使您更新现有脚本或创建新脚本。如果不及时更新它们，您可能会错误地转换和映射数据，并无意中造成数据质量问题。

为了防止这种情况，组织需要不断检查他们的 ETL 验证脚本是否已经过时，然后让他们的数据工程师花费数小时编写和重写重复的 ETL 验证脚本。这需要大量持续的工程时间和精力。它使您的数据工程师远离更有价值的活动，例如为业务构建新的解决方案。

此外，当您的数据工程师离开组织时，他们会随身携带有关您的 ETL 验证脚本的特定知识。这为每个替换数据工程师创造了一个陡峭的学习曲线。

为了应对当今快速增长、不断变化的数据环境，数据运营团队需要一个现代平台，该平台可以利用机器学习来自动监控所需规模的数据质量。

snowflake：“它只是有效”的规模和敏捷性

Snowflake 是当今最流行的云数据仓库之一。在短短十年内，该公司已发展到近 6,000 家企业客户和 12 亿美元的年收入，是上一年的两倍多。

与其他云数据仓库（包括 Databricks、Amazon RedShift 和 Google BigQuery）一样，Snowflake 拥有低启动成本、不断创新和“它只是工作”的可管理性的有吸引力的组合。从其惊人的增长和热情的客户来看，100% 的人推荐 Snowflake，Snowflake 在这些功能上的表现优于其竞争对手，尤其是在这两个领域：

1. 近乎零管理的高可用性。以最少的操作麻烦保证正常运行时间，因此您不需要庞大的 DBA、数据工程师等团队。

2. 即时部署的无限基础设施。虽然大多数公共云数据库在架构级别上将计算与存储分开，以使它们能够独立增长或缩小，但 Snowflake 具有特别的弹性，支持即时、自动扩展和缩减，可以顺利处理计划内或计划外的摄取数据或分析突发工作。无需订购硬件、配置集群或获得数据工程师的批准。

意外后果法则

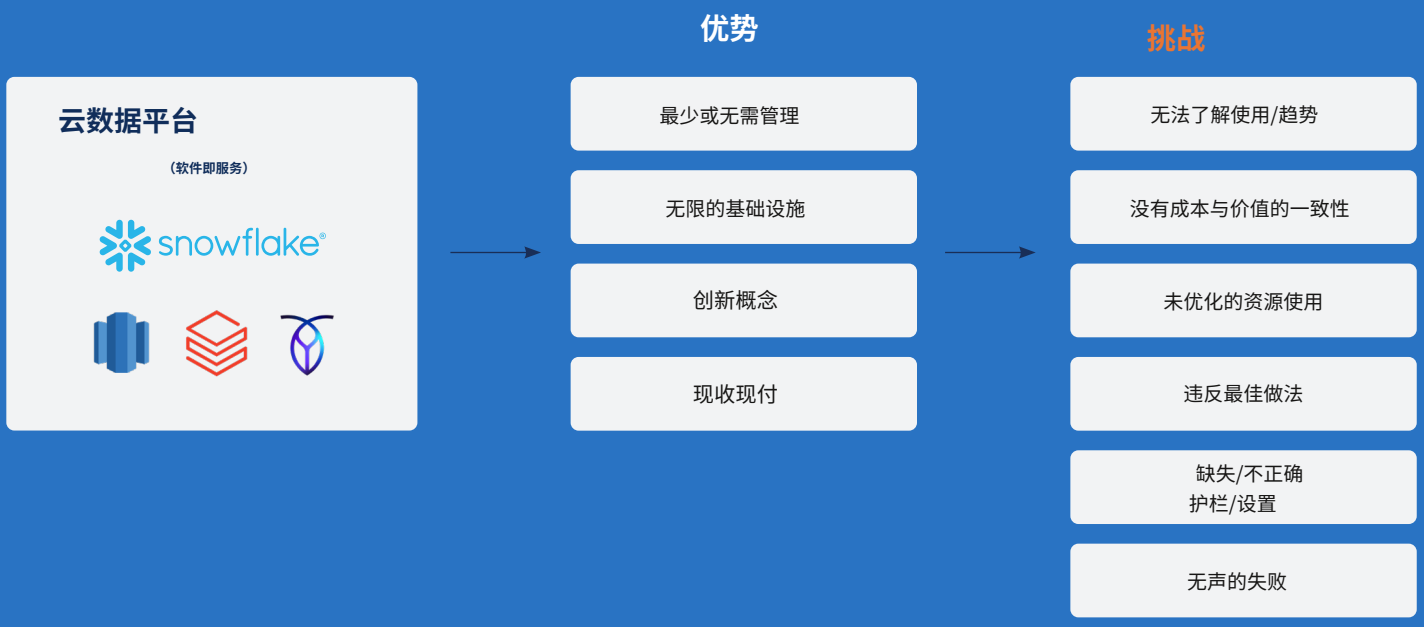
没有公司会拒绝如此简单的敏捷性。同时，运营和扩展 Snowflake 是一个梦想，它让许多公司忽略了一些重要的最佳实践。

其他人试图通过从其本地数据库和数据仓库中导入遗留数据质量流程来偷工减料，这些流程通常不适合并且会失败。

如果没有一个强大的数据运营团队遵循一套现代的最佳实践，这可能会导致混乱：数据存储库和数据管道的不受控制的增长以及没有适当的沿袭和跟踪的数据的猖獗复制。

数据错误悄悄潜入并在未经检查的情况下自行复合。数据管道失败，分析仪表板产生错误的结果。如果没有很强的可观察性，这些问题可能会持续数周或数月，然后用户才会注意到并抱怨。到那时，可能有许多 TB 或 PB 的数据需要重新扫描以查找数据质量问题并在整个组织中回填，具体取决于依赖项的数量和数据管道的复杂程度。您的snowflake数据云已变成数据沼泽。而snowflake的low ops也正式进入了存在的境界 *太多的好事。*

云数据平台：好处和挑战



但肯定还有像您这样的其他 Snowflake 用户优先监控和保持数据质量吗？是的，但事实证明这在 Snowflake 中很难做到。

Snowflake 的主要管理界面是一个名为 SnowSight 的 Web 仪表板。借助 SnowSight，用户可以监控查询性能和复制历史记录、创建和管理数据库和仓库等等。与其他数据库管理控制台相比，可见性和控制力有限。用户甚至无法配置 Snowsight 来推送实时警报或状态报告。对于习惯于对其数据进行持续可见性和控制的数据工程师来说，这可能会令人不安。

如今，大多数 Snowflake 用户使用 ETL 脚本来验证和清理传入的数据，因为它是由 Snowflake 处理和存储的。如果他们不使用 SnowSight，他们倾向于使用基于 SQL 的报告应用程序，例如 Tableau、Looker 和 Microsoft Power BI，它们可以摄取和显示 Snowflake 公开的操作数据。然而，这些可视化应用程序的基于批处理的设计使其最适合执行报告，而不是实时可观察性或日常管理。

与 SnowSight 一样，它们缺乏数据工程师监控、修复和预防一般数据问题所需的警报、预测能力和微调控制。特别是在数据质量方面，Tableau 及其同类产品缺乏开箱即用的模板和记分卡，使他们能够摄取、分析和显示 ETL 脚本生成的数据质量指标。**这意味着这些流行的报告工具完全不适合作为 Snowflake 操作仪表板，尤其是用于监控和优化您的数据质量。**

这是一个具体的例子，说明 Snowflake 善意的低操作尝试如何无意中造成数据质量问题。Snowflake 的数据仓库不需要用户管理分区或索引。相反，Snowflake 会自动将大表划分为微分区，并计算每列数据中包含的值范围的统计信息。这些统计信息有助于确定运行查询所需的数据子集，从而加快查询速度。

问题？从传统的分区和索引数据库迁移到 Snowflake 的数据需要在加载时进行转换，这可能会产生数据和架构错误。即使是一个小问题，例如 Snowflake SQL 代码的区分大小写，也可能导致应用程序和数据管道损坏。而且这些不太可能被 Snowflake 标记，也不太可能被您的数据工程师注意到。

依赖于 Snowflake 的指标，传统的 MDM 和数据治理工具无法捕捉 SQL 语法错误和其他更微妙的数据质量问题。同时，Snowflake Data Profiler 等数据分析工具只能绘制数据的高级概览，不能执行找出数据质量问题的实际检查。对于 Looker 等报告工具，他们只能在单个时间点（例如在摄取数据时）抽查不一致并测试数据质量。如果没有持续的数据质量验证和测试，他们不会注意到以后出现数据错误。

通过数据可观测性解决数据质量问题

Snowflake 用户正在转向现代数据可观测性应用程序，以帮助他们自动执行所需的连续数据验证和测试，从而在整个组织范围内建立对其数据的信任。

(>)但是，并非所有数据可观测性应用程序都是一样的。

(>)有些人只关注一个维度——数据质量——但缺乏对数据性能的任何洞察，也就是计算可观察性或如何优化数据的性价比。，

(>)其他人过于依赖 Snowflake 提供的指标和元数据，从而限制了他们的见解范围和预测能力。

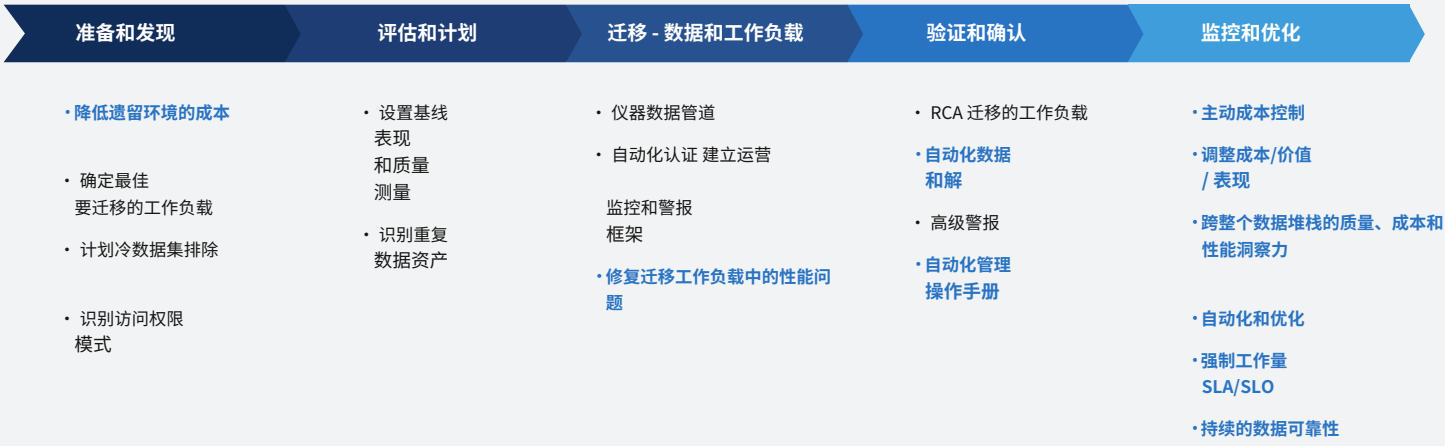
(>)还有一些只提供宏观层面的视图，而不是能够更准确和更准确地对数据进行切片和切块的能力。经济高效地确定数据错误的根本原因。

(>)最后，有些实际上是应用程序性能管理 (APM) 解决方案——想想 DataDog 或 New Relic——变相，试图将他们的应用程序级可观察性作为真正的数据可观测性（嗯，不是这样）。

公司需要的是一个多维云数据可观测性平台，该平台利用 Snowflake 的指标，并围绕 Snowflake 数据质量不断收集自己的统计数据和元数据。

然后，它将这些数据组合在一起，以生成自己的原始数据质量概况和见解。它使用这些更复杂的基线来自动分析和验证您的数据，因为它被摄取到 Snowflake 中。它会继续不断地验证和测试您的数据，以说明您的数据如何演变以及您的业务需求如何变化。

数据迁移的数据可观测性



HK-Acceldata 提供了这样一个多维数据可观测性平台，可以使解决数据质量成为现实的目标，而不是一个不可行的天上掉馅饼的目标。

在接下来的页面中，我们列出了实现这一目标的七种方法：

acceldata

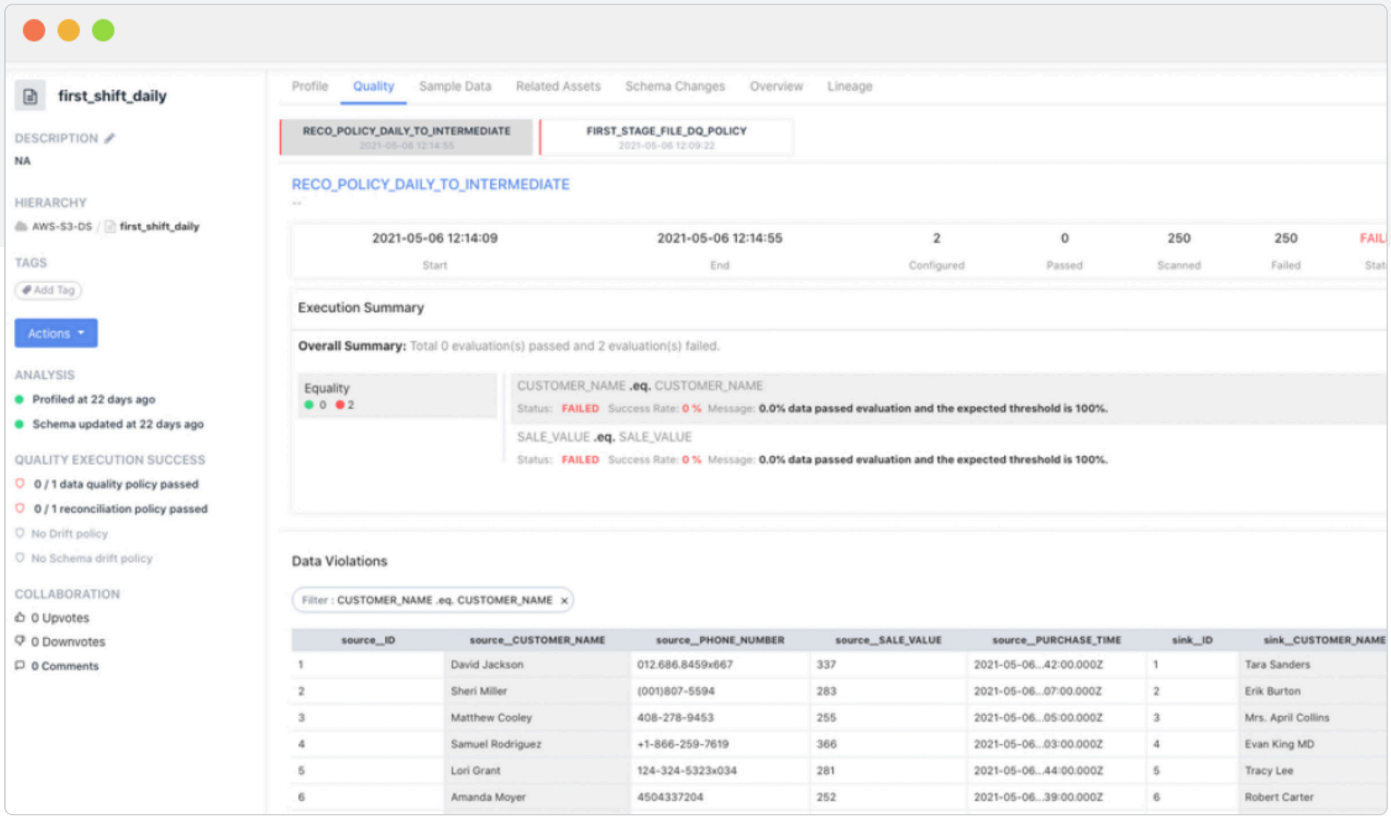
1. 显著简化的数据迁移

在迁移到 Snowflake 的每个阶段，HK-Acceldata 都会帮助自动化步骤，以帮助您最大限度地提高数据质量。

概念证明：HK-Acceldata Data Observability Cloud 可帮助您确定要迁移的基本工作负载、可以排除哪些未使用的数据集以及必须在 Snowflake 中重建哪些数据管道。

准备：HK-Acceldata 会自动创建数据目录，以便您识别可以从迁移中排除的重复数据资产。它还使用最佳实践配置您的 Snowflake 帐户和数据布局，因此您的 Snowflake 数据云是安全、高性能且经济高效的。

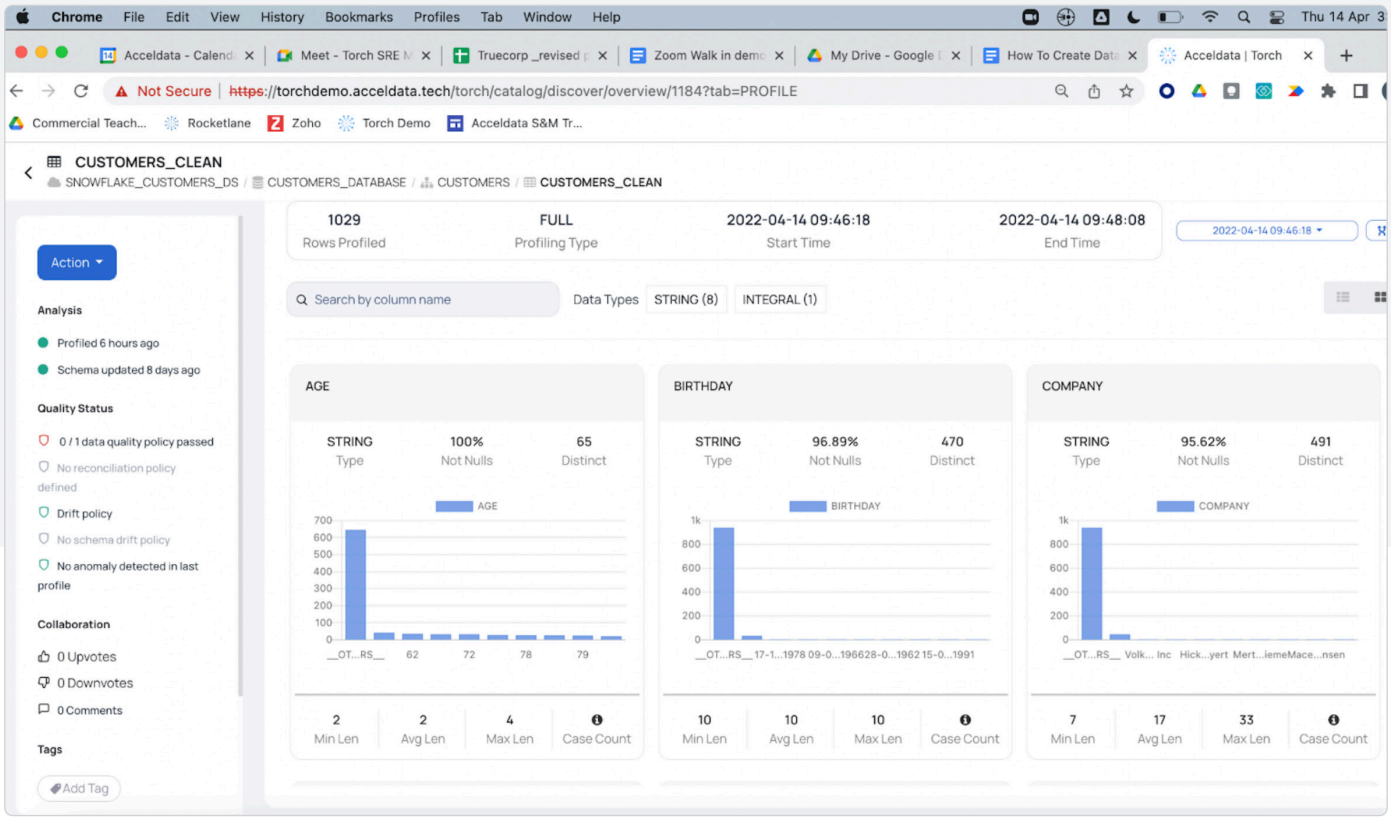
摄取：无论您使用 Snowpipe、COPY 还是其他途径，HK-Acceldata 都可以深入了解数据摄取过程。HK-Acceldata 还通过比较源数据集和目标数据集来验证数据。它还对未按预期工作的迁移工作负载进行根本原因分析 (RCA)。



2. snowflake数据云的自动分析

HK-Acceldata Torch 是我们平台的数据可靠性层，它还会自动发现您的所有数据集并创建所有数据的配置文件，包括它们的结构、元数据和关系，包括依赖关系和沿袭。

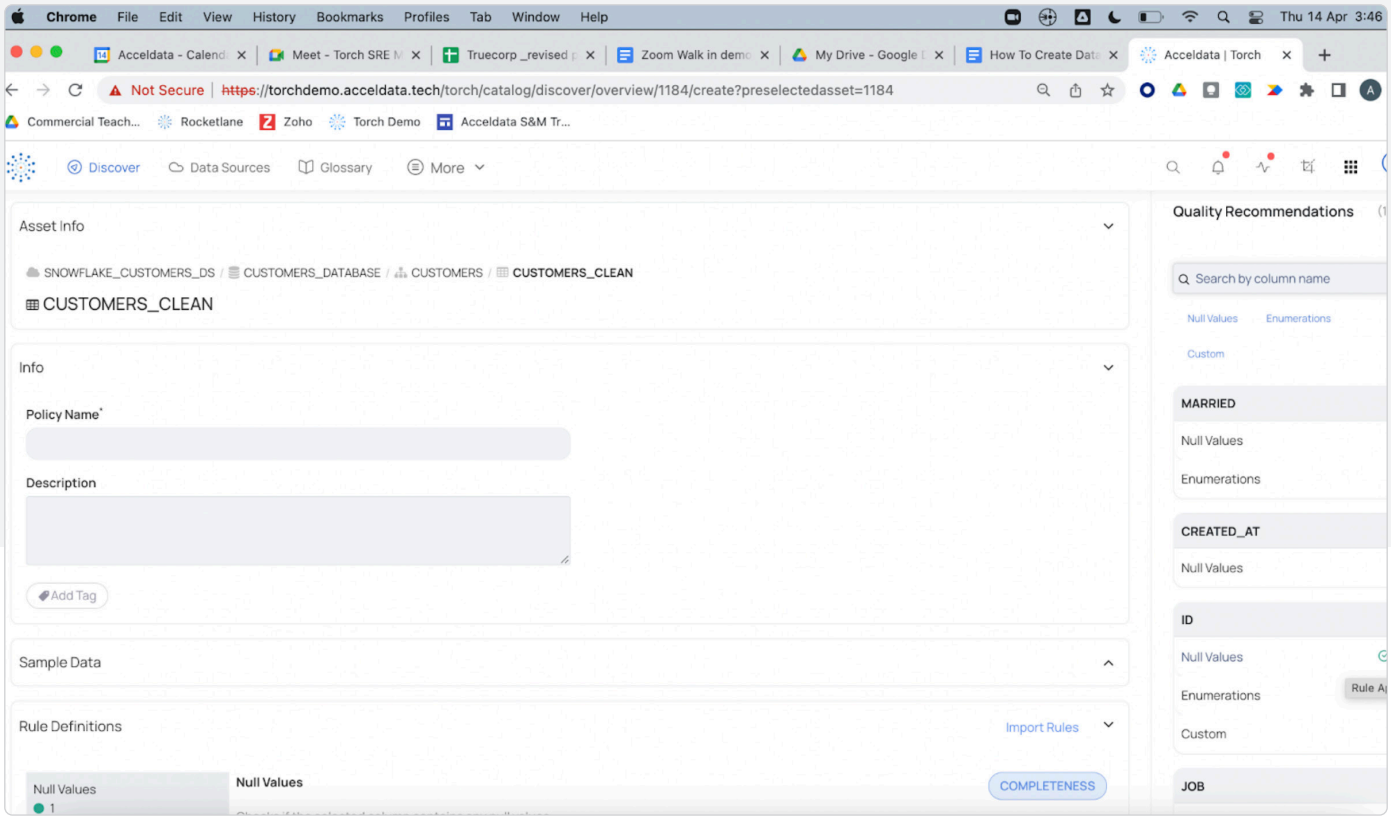
这是关键。没有这种背景，几乎不可能创建数据质量规则。HK-Acceldata Torch 中直观的交互式图表和图形为您的数据质量提供了急需的洞察力。



3. ML 支持的数据质量建议

Torch走得更远。分析您的数据后，HK-Acceldata Torch 开始提供基于 ML 的建议，以简化数据质量策略和规则的创建。

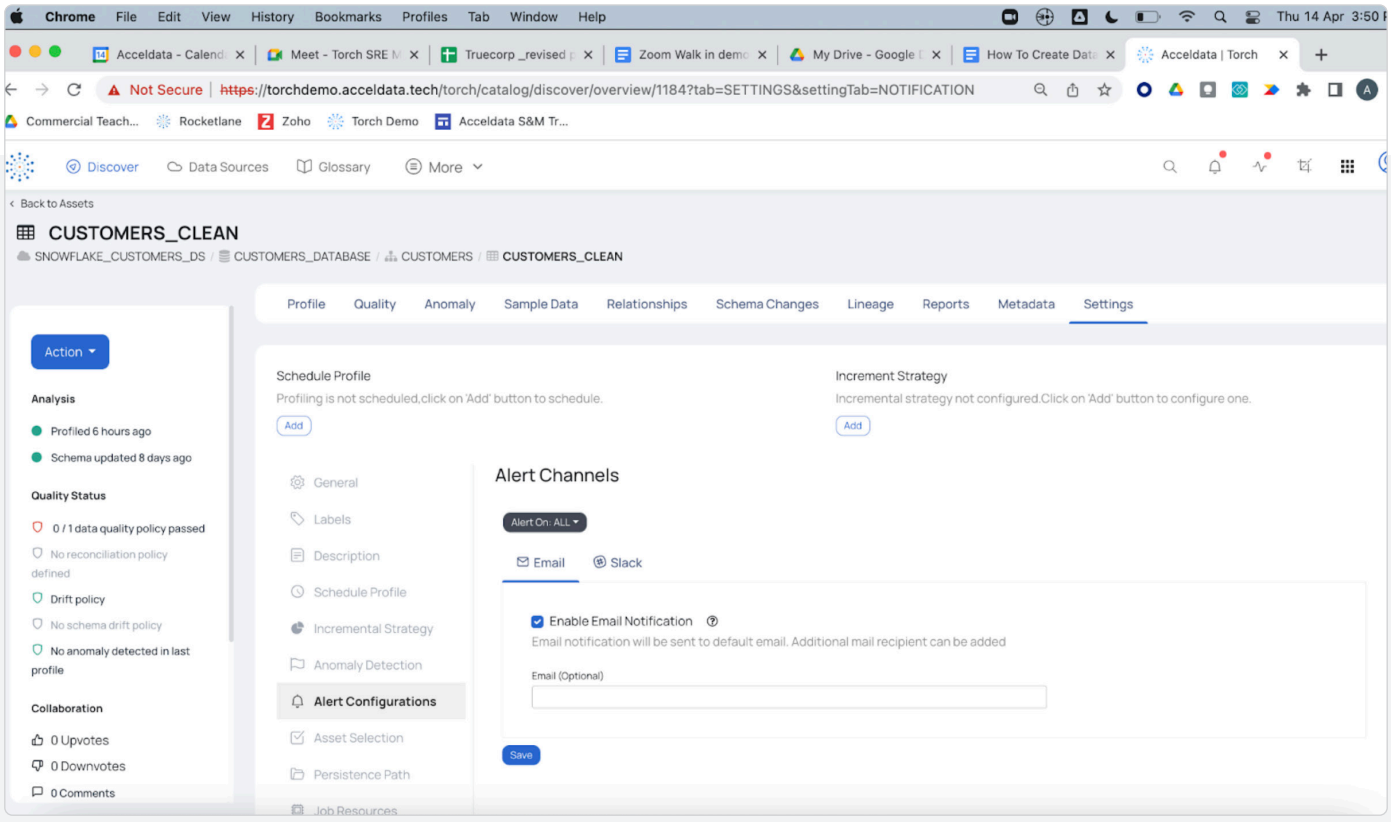
在以下示例中，HK-Acceldata Torch 识别出列中的数据应该是二进制的（“yes”或“no”）并且没有空值。只需单击推荐的规则即可将其添加到您的数据质量策略中。



空值只是一个例子。HK-Acceldata Torch 可以提出许多其他数据质量建议，包括：

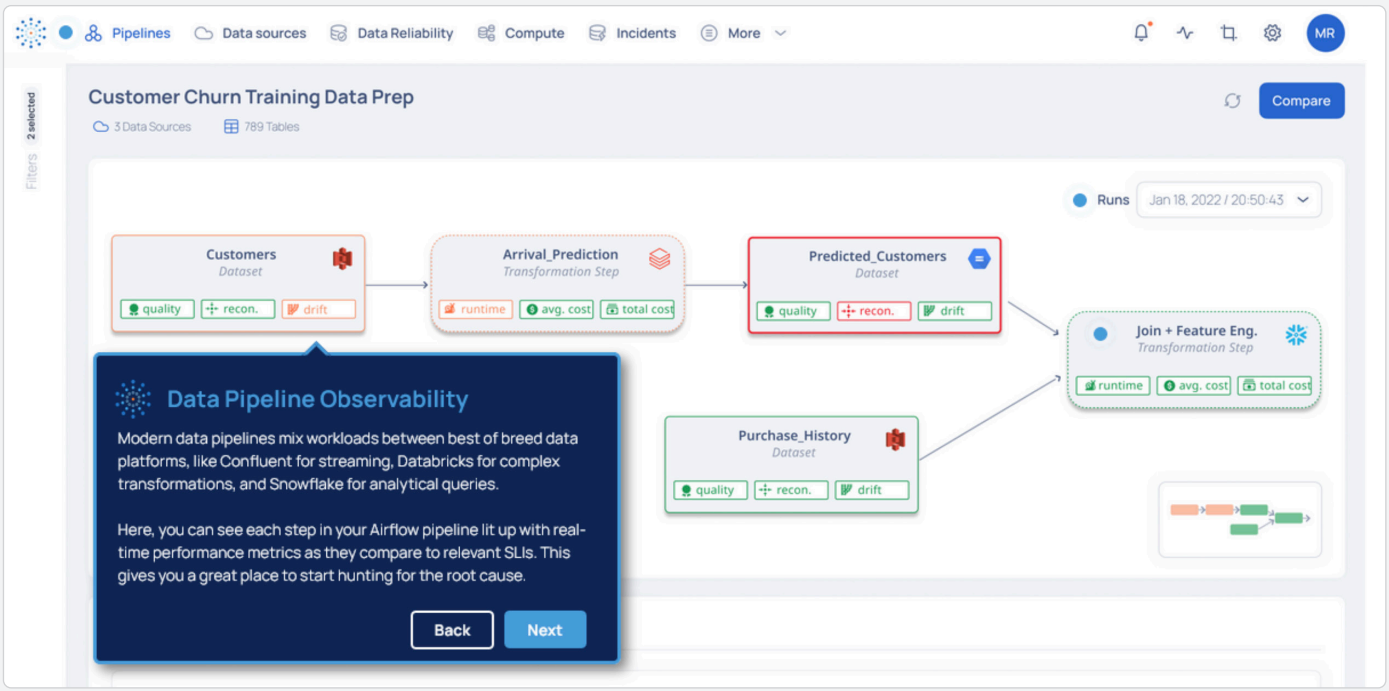
(>)枚举检查
(>)范围验证
(>)重复检查
(>)模式检查
(>)独特性
(>)自定义规则（公式）
(>)模式匹配

过去需要数小时的艰苦努力，现在只需单击几下即可在几分钟内完成。配置运行数据质量规则的计划也很简单。查看、编辑和删除您的数据质量策略也是如此。当规则执行时，您的数据运营团队可以收到电子邮件或 Slack 通知，以便您随时了解最新情况。



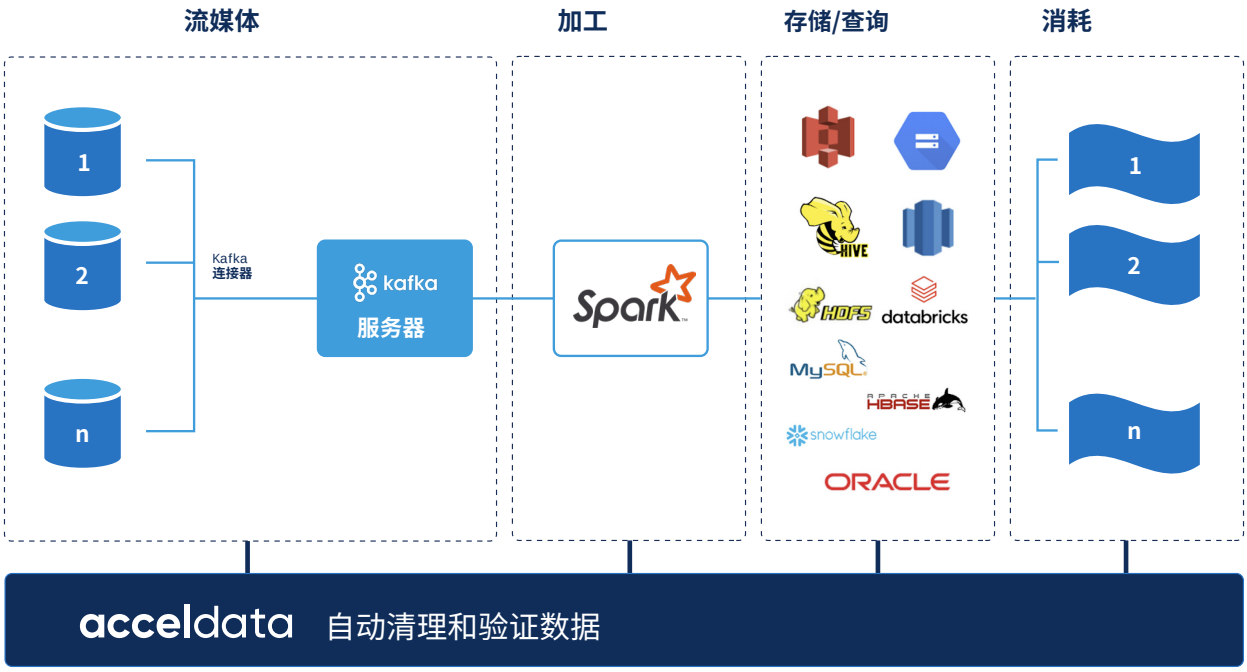
4. 持续的数据质量监控

如果没有持续的数据质量监控，设置基线和初始规则将毫无用处。HK-Acceldata 为您的整个数据管道的整个生命周期提供一个统一的视图。Torch 将持续监控和测量您的snowflake数据云的数据准确性、完整性、有效性、唯一性、及时性、模式/模型漂移和其他质量特征。即使数据通过不同的技术多次转换，这也有助于您保持可靠性。



HK-Acceldata Torch 还允许数据团队轻松检测数据结构的意外更改，或所谓的“模式漂移”，这可能会破坏数据管道或产生数据质量问题。Torch 可以检测数据中的异常和趋势（“数据漂移”），这些异常和趋势可能会通过数据质量检查，但仍然是一个问题。例如，可以检测到丢失的整个记录 或记录组。数据漂移可能需要重新训练 ML 模型以保持其准确性。此外，利益相关者可能希望收到有关数据趋势和异常的警报，这些趋势和异常代表了现有报告和仪表板上可能不可见的商业机会或威胁。Torch 还有助于协调从源到目标的数据，以确保数据保真度。这些功能共同帮助您避免数据管道损坏、数据质量差以及 ML 和 AI 模型不准确。

Torch 还可以自动对未分类的原始数据进行分类、聚类和关联。这有助于数据团队理解大型数据集。这为每个数据记录如何与其他记录关联提供了上下文。



5. 精细且具有成本效益的数据质量分析

我们听取了 Snowflake 用户的反馈，他们说被迫分析整个庞大的数据库表通常是浪费、太昂贵且速度慢。使用 Torch 强大而简单的 UI，数据工程师可以定义他们想要探索的数据质量或他们希望应用数据质量规则的数据段。

通过分段分析，用户可以将数据质量检查限制在可能有问题的行或列、新数据或高优先级数据。Torch 甚至允许用户通过字段本身的内容来限制数据质量分析，例如 N>1000 的值，或者文本仅为“女性”时。Torch 还允许您比较不同细分市场的数据质量和健康状况。

6. 自动查找异常和根本原因问题

HK-Acceldata Torch 使用机器学习来分析 CPU、内存、成本和计算资源的历史趋势，并发现可能表明数据质量问题的异常。它通过对可以验证或不验证的数据做出断言来持续运行测试。

Torch 还可以通过比较应用程序日志、查询运行时或队列利用率统计信息来自动识别意外行为的根本原因。这让团队可以避免手动筛选大型数据集来调试数据质量问题。它还可以帮助他们快速确定哪些下游数据用户受到的影响最大。除了为您的团队节省宝贵的时间外，它还可以降低存储和处理成本，同时提高性能。

7. 自动清理和验证实时数据流

公司越来越多地部署实时客户个性化、数据驱动的物流、欺诈检测和其他内部操作、实时用户排行榜等。此类颠覆性用例依赖于事件和消息流形式的实时数据、变更数据捕获(CDC) 更新等。为了使这些数据对实时分析有用，首先需要立即对其进行清理和验证。

HK-Acceldata 可帮助您监控和自动化实时数据源的清理和验证，例如连接到snowflake数据云的 Apache Kafka。HK-Acceldata 首先分析存储在您的 Kafka 集群中的数据并监控事件以获得更快的吞吐量和更好的稳定性。

然后，HK-Acceldata 会自动实时标记不完整、不正确和不准确的数据，而无需您的团队进行手动干预。这可以保持数据流动，并将数据停机时间降至最低。

HK-Acceldata 是灵活的。除了 Kafka，HK-Acceldata 还集成了 Spark 等处理引擎，以及 Amazon S3、Hive、HBase、Redshift 和 Databricks 等其他云存储和查询平台，以及 MySQL、PostgreSQL 和 Oracle 等遗留系统。

自动清理和验证您的实时数据流可以让您的数据团队腾出时间进行创新。它只是 HK-Acceldata 统一的多维数据可观测性平台的七个功

能
之一，该平台可在数据在整个生命周期中传输和转换时提供简单、完整的数据可追溯性。
其他解决方案缺乏多层可见性或提供不完整的视图。这迫使您将各个解决方案拼凑在一起，每个解决方案都提供不同的、不完整的视图。这会造成数据碎片化，因为数据团队无法再端到端地观察数据。而这会导致数据管道中断、莫名其妙的数据质量问题以及意外的数据中断，进而需要数据团队手动调试这些问题。

入门

低操作≠没有操作。像您这样的公司意识到，即使使用 Snowflake，您也必须投入时间和精力来构建一个连续的数据质量周期，以便测试、验证和改善出现的数据错误。

帮助您创建和自动化此连续数据质量周期的最佳技术合作伙伴是像 HK-Acceldata 这样的数据可观测性平台，它在您的系统中的整个过程中提供数据的单一统一视图。

详细了解 HK-Acceldata **如何通过对性能、质量、成本等方面的深入了解，帮助您最大限度地提高 Snowflake 投资的回报。**