

HK-Acceldata 解决方案指南

1 HK-Acceldata Pulse 解决方案指南

HK-Acceldata Pulse

观察混合数据环境以优化可靠性、性能和成本

随着数据驱动型组织从比以往更多的来源访问数据，IT部门正面临着管理、处理和存储这些数据的方式和地点的转变。如今，IT不仅管理数据库以运行关键任务应用程序和数据仓库以促进分析。

它还监督用于数据探索的大数据湖、用于动态数据的事件流处理器以及用于灵活数据访问的云数据管理解决方案。

随着数据在复杂的数据环境中从源移动到目标，可能会发生影响可靠服务和及时性能的事件。当数据交付中断时，业务活动就会放缓，甚至更糟糕的是，会陷入停顿。

为确保在需要时将准确的数据交付给现代数据应用程序，数据团队需要一种数据可观察性工具，该工具可提供跨混合数据环境的全面洞察力。HK-Acceldata Pulse 可以提供帮助。

什么是 HK-Acceldata Pulse?

HK-Acceldata Pulse 是一种数据可观察性工具，可全面了解混合数据环境和由多种技术拼凑而成的管道。它在数据通过本地或云中的连接数据源移动时观察数据，以根据业务定义的指标监控数据处理值，以确保底层系统的可靠性和性能。借助人工智能和

机器学习，Pulse 将来自基础设施、应用程序和数据层的信号关联起来，以随着时间的推移开发上下文。凭借内置的专业知识，Pulse 能够自动执行操作以解决问题、减少重复发生并提高生产力。Pulse 从单一管理平台将混合系统中的问题整合到一个整体视图中，使数据团队能够快速响应问题、预测和预防未来问题、优化现有资源、快速扩展并有效管理成本。

提高对混合数据环境的可见性

HK-Acceldata Pulse 使 IT 专业人员 and 数据工程师能够通过连接到数据源、定义要观察的指标和阈值以及从可定制的仪表板监控结果，从而在混合数据环境中获得更大的可见性。

监控数据处理性能

HK-Acceldata Pulse 为 Spark、Kafka、Nifi、Impala、Hive、HBase、Hadoop 等开源技术提供连接器。通过监控每个作业的基础架构容量、处理能力和资源利用率，获得整个堆栈的数据性能的统一视图，无论工作负载在何处运行（本地或云中）。有关连接器的完整列表，请参阅“HK-Acceldata Pulse 集成”。

优化数据性能以满足业务需求

HK-Acceldata Pulse 帮助 IT 满足并超越业务需求。对于每个数据源，数据工程师可以在基础设施、数据和应用层设置数据性能指标或阈值。例如，容量、CPU 或内存使用等基础设施指标有助于衡量稳定性和规模。表、文件或查询的数量或执行查询的用户数量等数据指标可以帮助数据工程师提高性能和规模。

服务状态、资源使用或错误率等应用程序指标可以帮助确定数据是否适合应用程序。Pulse 提供有关堆栈中每个元素如何运行的持续智能，并帮助数据团队发现组件如何在分布式数据环境中相互依赖和交互。

查看丰富的仪表板以确保业务成果

随着数据流经复杂的管道，Pulse 根据用户定义的指标监控连接的数据源，以满足服务水平协议(SLA)。它以丰富的数据可视化形式将结果交付给可定制的仪表板，使数据团队能够从稳定性、可扩展性和响应性方面了解系统运行状况和性能。

确保数据可用性、可靠性和性能

HK-Acceldata Pulse 通过在出现问题时通知 IT 和数据团队、使用热图突出显示热点、发现模式、计算随时间变化的差异水平以及提供 AI 支持的建议来解决问题来帮助减少数据停机时间。通过更清楚地了解影响可靠性和性能的因素，数据团队可以花更少的

时间进行故障排除，而将更多的时间用于优化资源和扩展系统以改善利益相关者的成果。

出现问题时迅速得到通知

当出现异常或超过阈值时，HK-Acceldata Pulse 会提供开箱即用的警报。数据工程师还可以通过定义基于度量的条件或阈值、严重性级别（即低、中、高或严重）以及在值超出指定条件时执行的操作来创建自定义警报。

如果针对特定阈值继续出现警报，则会为警报创建事件。事件指定警报发生的次数和持续的秒数。警报或事件都可以触发通知或自动操作。可以通过电子邮件、Slack、webhook 等方式发送通知。自动操作是一组预定义的活动，将自动采取以解决警报。

HK-Acceldata Pulse 支持 IT 专业人员和数据工程师通过连接数据源、定义来自可定制仪表板的指标、观察阈值和监控结果。

快速识别和解决热点

HK-Acceldata Pulse 可帮助数据工程师通过可视化指标（例如颜色编码的热图）识别服务器或内存使用热点。例如，图 1 显示了内存和 VCore 使用量的峰值，这似乎是由营销部门推动的。借助此信息，IT 可以根据 SLO 衡量实际使用情况，以验证预期行为或跟踪异常情况。Pulse 提供补救建议以加快问题的解决。

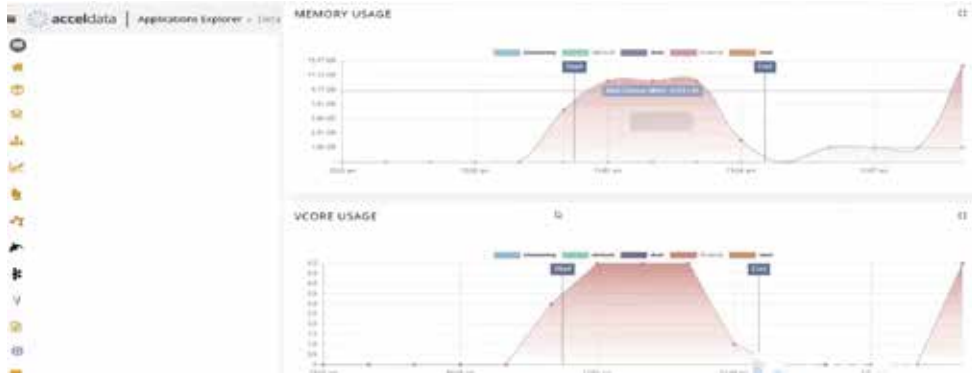


图1: 使用颜色编码的热图识别和解决热点

发现根本原因

HK-Acceldata Pulse 可帮助数据工程师执行根本原因分析，以确定应用程序和查询性能。随着数据在分布式系统中移动，Pulse 会收集有关进程的信息，例如运行时和配置模式，并将相应的事件存储在应用程序日志中。数据工程师可以搜索日志、跟踪错误并关联不同时间间隔的运行，以确定问题的根本原因。例如，图 2 以黄色突出显示配置差异，使数据工程师能够深入了解为什么一项作业成功运行而另一项作业失败。

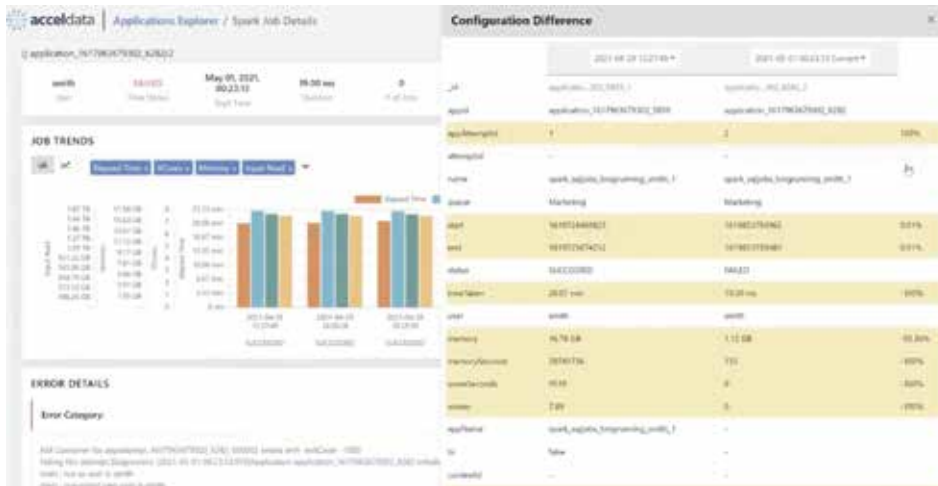


图2: 跟踪和比较作业以执行根本原因分析

在问题发生之前发现模式并预测问题

HK-Acceldata Pulse 不仅仅是在数据流经数据管道时监控数据。它随着时间的推移观察、学习和发现模式，以帮助数据工程师在可靠性或性能问题出现之前预测和解决它们。特别是，Pulse 计算方差水平以确定工作负载随时间的变化程度。数据工程师可以每小时、每天、每周或每月分析和比较特定工作的数据处理性能，如图 3 所示。

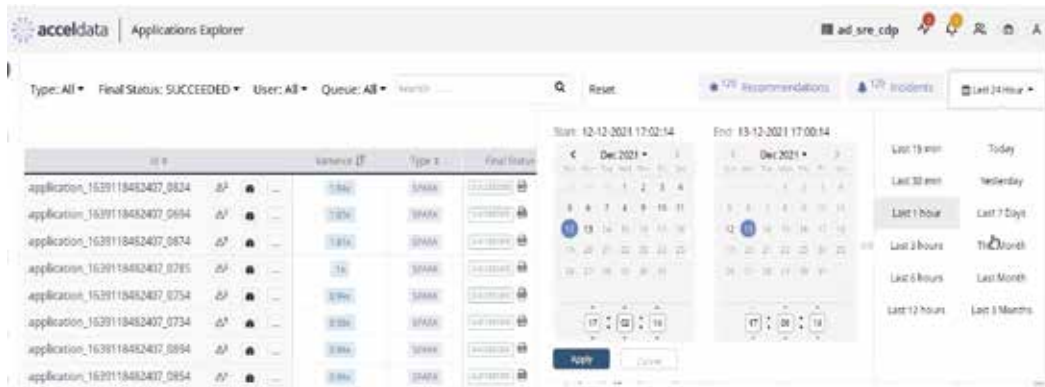


图 3: 比较 Spark 运行时进程中的方差水平以执行趋势分析

在此示例中，第一个作业显示 1.04 倍的方差水平，表明该作业的完成时间比正常时间长 39%。虽然工作成功完成，但这种差异水平是不稳定的预警信号。通过深入研究工作，数据工程师可以评估磁盘容量、内存使用或资源消耗，以预测和防止未来潜在的 SLA 承诺违规。

自动化预防性维护、性能调整和问题修复

HK-Acceldata Pulse 为不同的技术提供自动操作，并使数据工程师能够创建新的操作来自动化需要频繁执行的工作或流程。可以计划、通过警报激活或使用工作流自动执行操作，以防止问题发生。例如，图 4 显示了用于管理 Hadoop 分布式文件系统性能的自动化操作列表。



图 4: 使用自动化操作来防止问题再次发生

通过基于 AI 的建议提高生产力

HK-Acceldata Pulse 提供基于 AI 的建议，帮助数据工程师快速解决问题并优化应用程序和分析工作负载的数据处理性能。它通过持续收集数据并关联有关服务器、数据、队列、应用程序、用户和其他资源的事件来理解和推荐解决方案来做到这一点。

例如，在观察和分析成功的 Spark 作业之后，Pulse 提供了一些建议，数据工程师可以使用这些建议来解决具有许多小任务、任务倾斜运行时和低最大内存的应用程序，如图 5 所示。

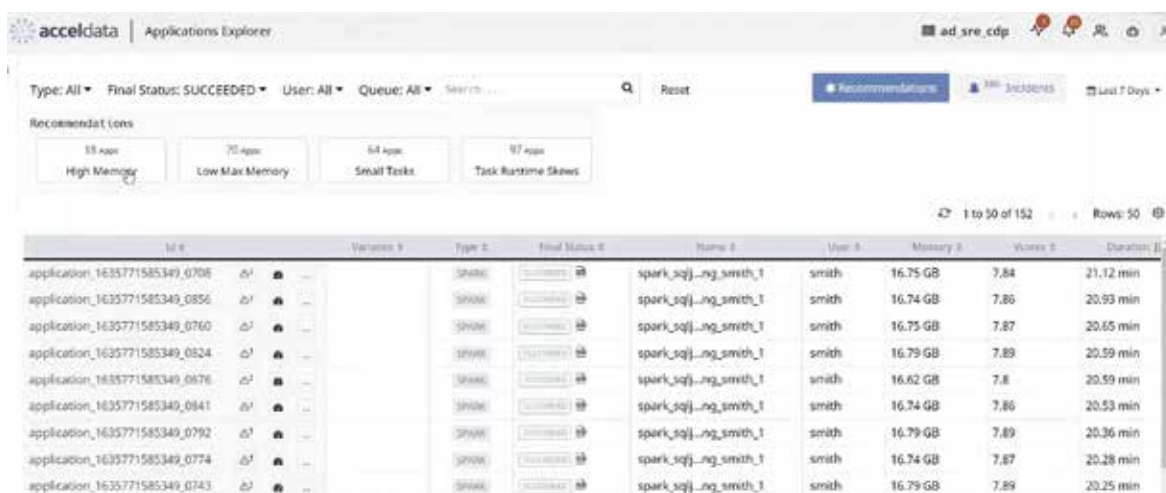


图 5: 使用颜色编码的热图识别和解决热点

减少瓶颈和不必要的开销

HK-Acceldata Pulse 允许数据工程师针对作业运行工作负载分析，以识别瓶颈和不必要的开销。工作负载分析可以识别影响处理的小任务、运行时偏差或不均匀的数据分布。例如，Pulse 突出显示了几个工作，这些工作正在运行大量的小任务（即 900、1、1），运行时间有偏差，如图 6 所示。有了这些信息，数据工程师可以改进应用程序的代码以减少数量小任务和改进并行处理。



图 6: 跟踪和比较作业以执行根本原因分析

运行仿真模型以满足服务水平协议 (SLA)

HK-Acceldata Pulse 使数据工程师能够运行仿真模型来评估满足服务水平协议 (SLA) 和控制成本所需的最少资源。通常，运行一次模拟以设置基线。随后的模拟用于优化资源。例如，图 7 显示，当分配了两个资源时，作业运行时间不到 15 分钟。当分配了四个资源时，作业将在不到 7 分钟内运行。但是，Pulse 显示，在分配 12 个资源之前，性能不会继续提高，而且提高幅度很小。因此，最具成本效益的解决方案是为该作业分配四个资源，保证 SLA 少于 7 分钟。

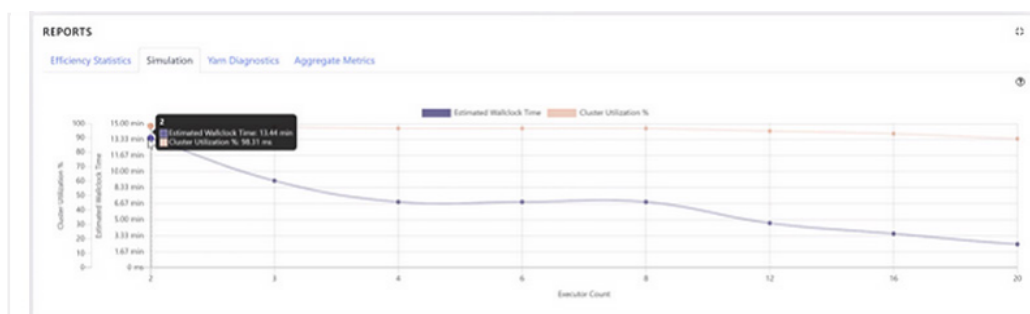


图 7: 运行模拟以评估理想配置

创建计费报告以控制成本

HK-Acceldata Pulse 在特定时间段内监控共享集群资源以创建计费报告。这些报告可以跟踪和衡量队列、服务、应用程序和用户的使用情况。有了这些信息，IT 组织可以更好地了解哪些因素在推动成本，并向业务部门收取消费费用。例如，图 8 显示了不同部门的队列容量利用率。



图 8: 创建计费报告以控制成本

概括

确保数据可靠性、提高性能、优化资源和降低成本是大多数数据驱动型组织的重中之重。然而，当数据分布在复杂且相互依赖的系统中时，实现这些目标并不容易。Accledata Pulse 帮助 IT 部门和数据团队超越简单的监控和故障排除，提供以下优势：

- > 提高分布式数据平台的可见性快速解决问题并减少平均识别时间 (MTTI) 和平均解决时间 (MTTR)
- > 快速解决问题并减少平均识别时间(MTTI) 和平均解决时间 (MTTR)
- > 提高现代数据应用程序的可靠性
- > 优化数据驱动企业的运营绩效
- > 快速有效地扩展以支持业务成长与创新
- > 提高数据投资回报
- > 提高数据团队的生产力并降低 IT 成本

随着复杂的企业数据系统以前所未有的速度部署到生产中，数据可观察性对于寻求将自己转变为能够在现代经济中竞争的数据驱动型企业的公司来说绝对至关重要。

获取演示

获取个性化演示 [请联系我们](#)。

2 HK-Acceldata Torch 解决方案指南

HK-Acceldata Torch

通过数据管道观察数据流以优化数据可靠性、质量和使用

数据驱动的组织在云端和 本地处理的数据比以往任何时候都多。在这种动态环境中，数据可能会变得支离破碎，从而难以获得企业中所有可用数据的完整和准确视图。

提供可靠且结构良好的数据的一种方法是通过数据管道自动移动和转换数据。现代数据管道可帮助组织快速连接到数据源，识别和提取相关数据，将其转换为可用形式，并将其加载到系统中以支持决策制定和分析。因为数据管道从不同的来源整合到一个共同的目的地，它们帮助组织获得企业数据的完整视图。

但是，数据管道可能很复杂。随着数据从源流向目标，它可能会经历多个处理步骤，将原始数据转换为可供应用程序或分析工具使用的数据。工作流定义了数据管道中作业的顺序以及它们之间的依赖关系，但是当发生错误时，可能很难发现和解决问题的初始根源。为了确保数据适合使用，数据管道需要持续监控以检查数据准确性和管道可靠性。HK-Acceldata Torch 可以提供帮助。

什么是HK-Acceldata Torch?

HK-Acceldata Torch 是一种数据可观测性解决方案，可提供对流经复杂数据管道的数据的可用性、可靠性和质量的可见性。当数据从源传输到目标时，Torch 会根据用户定义的策略监控数据，以识别数据和架构错误。它还监控数据管道本身的可靠性。通过机器学习，Torch 可以推断问题、自动发出警报并提供基于 AI 的建议，以帮助数据工程师快速解决问题。Torch 还监控数据沿袭，以帮助数据团队在问题影响用户、机器学习模型或应用程序工作流之前追踪根本原因并防止问题发生。最后，HK-Acceldata Torch 简化了数据管理，并实现了数据访问，促进了数据工程师、分析师和数据管理员之间的协作

借助跨数据管道的端到端可见性，Torch 可帮助数据团队更好地将其数据战略与业务需求保持一致。他们可以调查数据问题如何影响业务目标，隔离影响业务功能的错误，确定工作的优先级，并根据业务紧迫性和影响解决效率低下的问题。

获得连接数据资产的可见性

HK-Acceldata Torch 帮助数据团队监控数据源及其资产。它为数据库、数据仓库、数据湖、文件系统、分析工具、集成服务和 workflows 平台提供开箱即用的连接器。此外，数据工程师可以根据需要创建新的。有关连接器的完整列表，请参阅“HK-Acceldata Torch 集成”。

连接后，HK-Acceldata Torch 通过自动元数据分类、数据分析和数据编目简化数据管理，为组织提供所有可用数据的全面视图。它显示有关每个数据源的信息、跟踪数据资产并报告用于分析的数据质量，如图 1 所示。

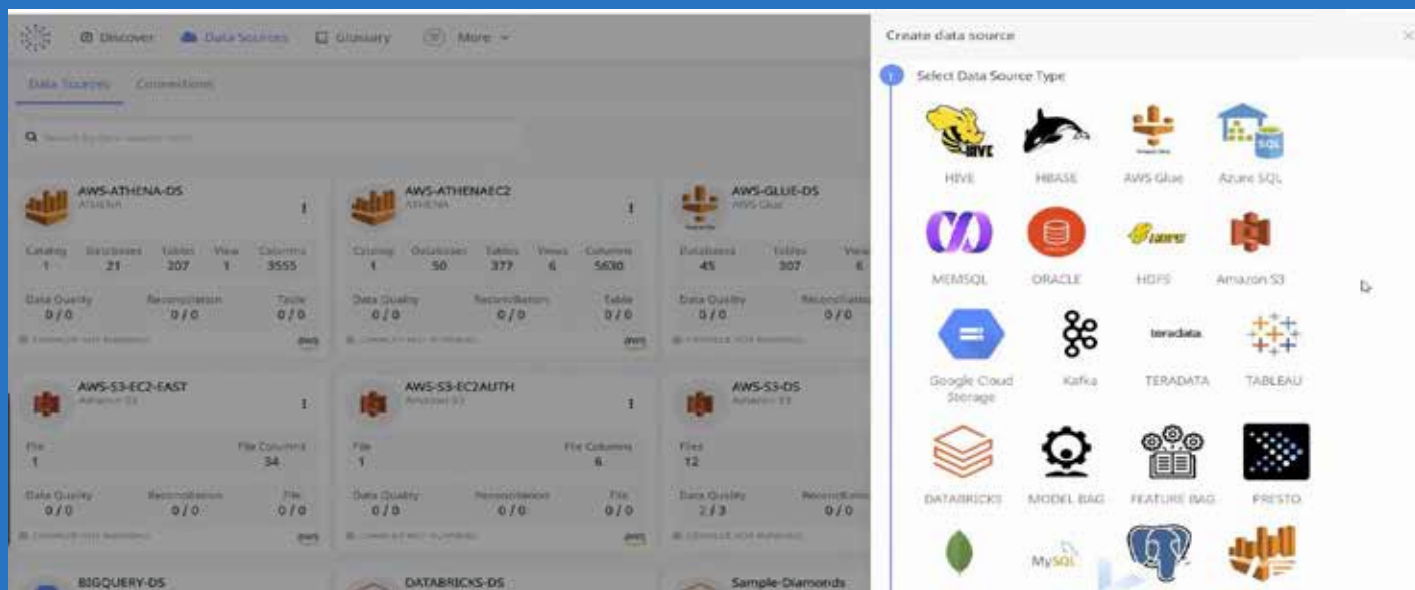


图 1: HK-Acceldata Torch / 数据源窗口

通过策略检查确保对数据的信任

HK-Acceldata Torch帮助数据团队提高对用于分析和机器学习的数据的信任。它根据用户定义的数据质量策略监控数据资产，协调、数据漂移和架构漂移。当出现问题时，Torch会提供自动警报和建议，使数据团队能够立即采取行动。

通过监控数据质量策略的合规性来确保数据完整性

HK-Acceldata Torch帮助数据团队创建分配给资产类型的数据质量策略。对于每个策略，数据工程师定义了质量规则，可以检查缺失数据、重复记录、不正确的格式、值或模式、无效标签、超出范围的值或业务规则。在数据流期间，Torch根据质量规则监控数据完整性、识别事件并显示仪表板以帮助数据团队快速解决问题。例如，图 2 显示了针对“Snowflake_Customers”表的策略检查失败，因为模式匹配，唯一性和空列规则被打破。

用户可以深入了解每个质量错误以获取更多信息或获得自动建议以快速解决问题。

使用协调策略提高数据流后的数据准确性

HK-Acceldata Torch利用协调策略确保数据按预期到达。对于每个协调策略，数据工程师会命名数据源和目标、要执行的比较类型以及要检查的比较，例如：

> Last name = last name (100% 平等阈值)

> First name = first name (75% 平等阈值)。

Torch在仪表板中显示结果，并可能会向人员发送警报，以便立即采取行动。例如，图 3 显示了两个表之间的违规列表，并显示协调策略失败，因为它不符合“First_Name”和“Married”的协调规则。

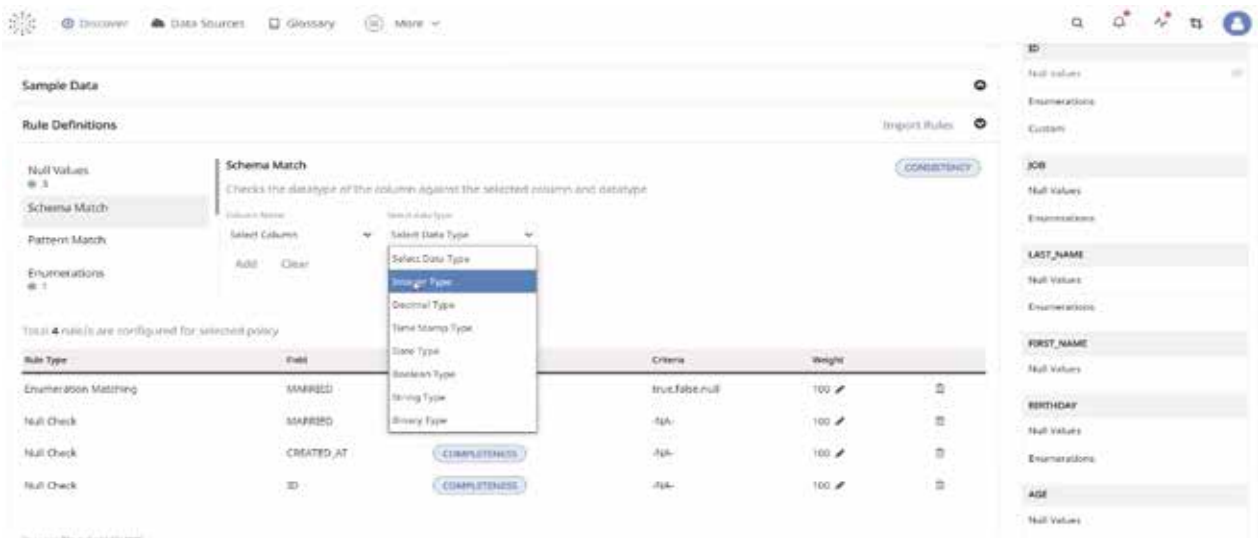


图 2: HK-Acceldata Torch / 策略 / 数据质量策略详情窗口

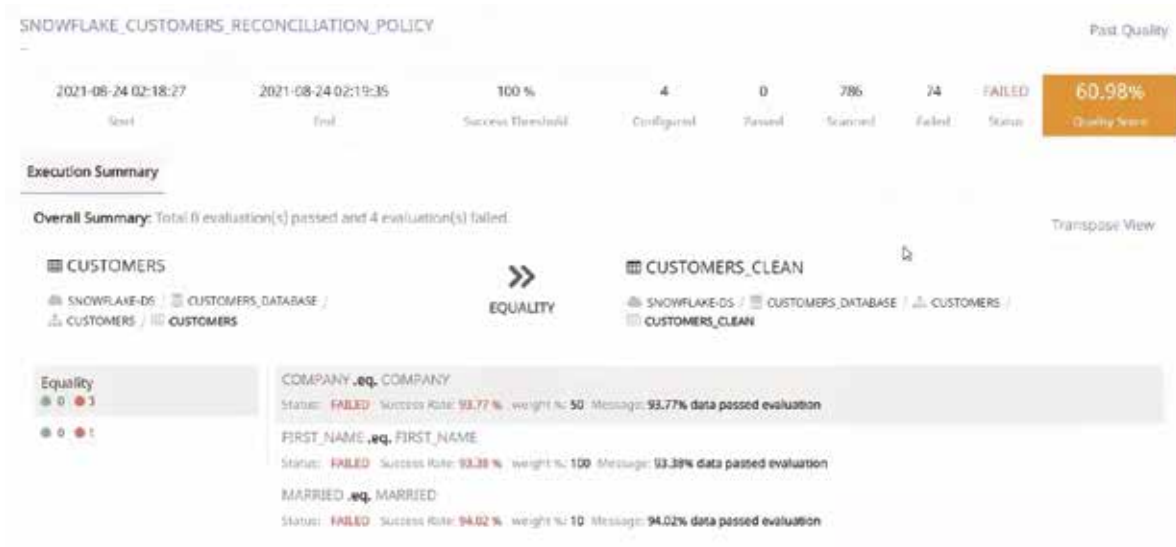


图 3: HK-Acceldata Torch / Policy / Reconciliation Policy Details 窗口

使用数据漂移策略减少数据错误

HK-Acceldata Torch 使数据工程师能够创建数据漂移策略，以减少数据流经管道时的意外更改。数据漂移策略根据容差阈值数据特征测量和验证数据，例如完整性、不同值、平均值、最小值、最大值、总和、标准差和前 10 个值。

在 Torch 中，每次分析资产时都会执行数据漂移策略。在数据分析期间，Torch 收集有关数据结构、部分如何相互关联以及单个记录中的错误的信息。Torch 还跟踪执行的每个配置文件。通过比较针对同一数据资产运行的两个配置文件之间的差异，数据工程师可以识别数据漂移错误何时首次出现。

使用模式漂移策略减少下游事件

HK-Acceldata Torch 帮助数据团队监控架构漂移。当数据源偏离用于定义其结构的原始模式时，就会发生这种情况。在 Torch 中，每当云或大数据爬虫爬取数据源时，都会执行架构漂移策略。在数据爬网期间，Torch 收集有关数据源的元数据（例如，表、列、字段和视图）并显示对架构所做的更改。当变化出乎意料时，数据工程师可以深入研究架构更改以了解发生了什么并在问题影响下游应用程序之前解决问题。

使用基于 AI 的建议和自动警报快速响应

随着 HK-Acceldata Torch 通过机器学习建立专业知识，它开始预测数据质量问题并提供基于 AI 的建议，数据团队可以应用这些建议来快速解决常见问题。这些基于 AI 的建议可以帮助消除手动步骤、提高准确性并提供更好的覆盖范围。

通过对数据管道的端到端可视性，
Torch 帮助数据团队更好地将其数据战略与业务需求相结合。

此外，HK-Acceldata Torch 在事件仪表板上显示所有策略检查的结果，如图 4 所示。成功的执行被标记为绿色。失败的执行被标记为红色。一旦事件得到解决，它们可以被标记为已解决。

HK-Acceldata Torch 提供对企业中使用的简单或复杂数据管道的更大可见性。一个成功的管道有效地移动数据，最大限度地减少任务之间的停顿，并保持流程运行。Torch 帮助数据工程师识别要观察的管道，使用 Apache Airflow 等工作流工具管理数据管道，在数据流经管道时分析数据沿袭，并在潜在资源或质量问题影响下游流程之前解决它们。

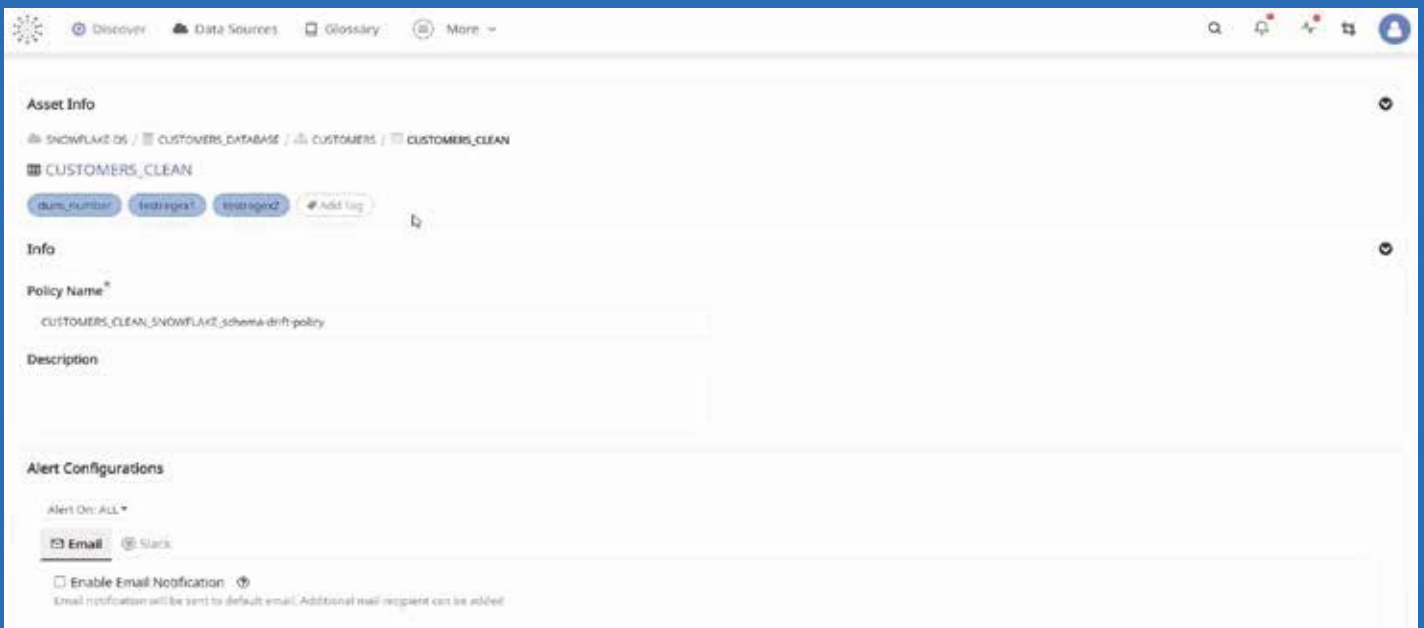


图 4: HK-Acceldata Torch / 资产信息窗口

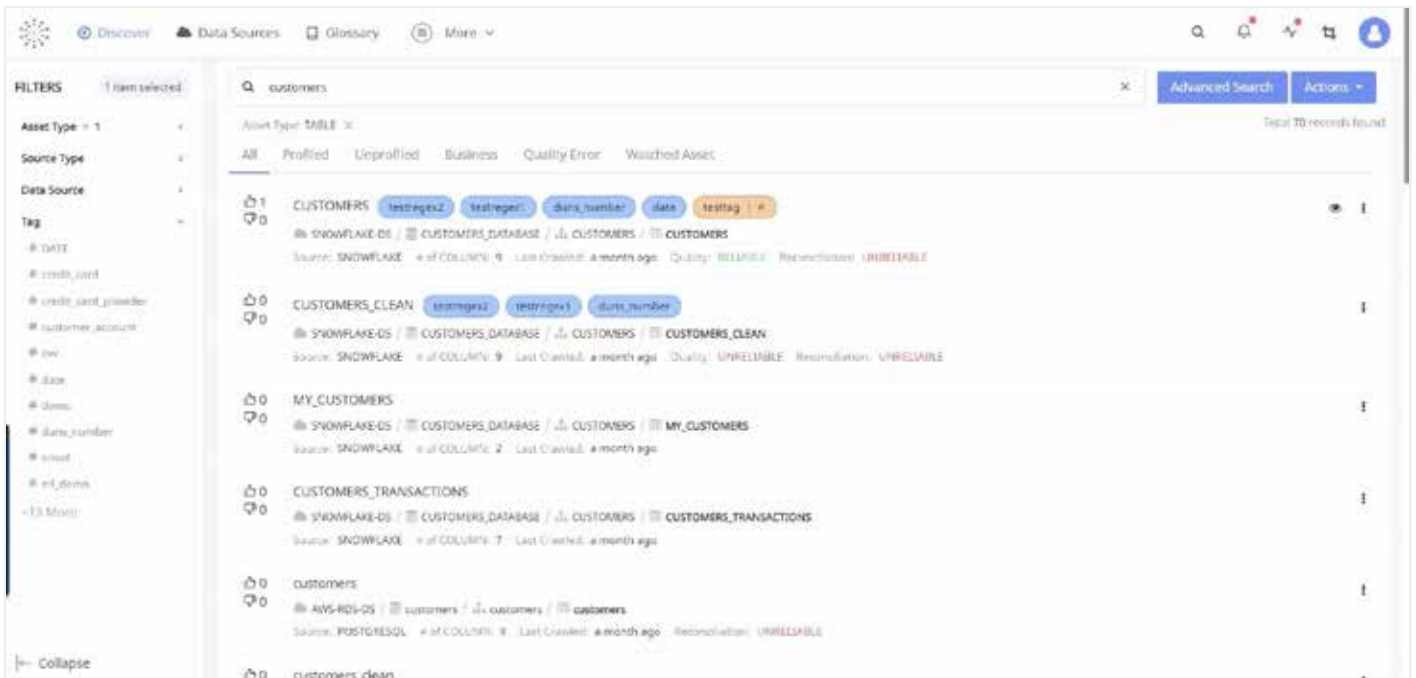


图 5: HK-Acceldata Torch/管道

确定要观察的数据管道

借助 HK-Acceldata Torch，数据工程师可以识别他们想要观察的每个数据管道的关键组件。提供了开箱即用的连接器，并且可以轻松添加新的连接器以识别：

> 来源：提取数据的地方（即 RDBMS、ERP、CRM、物联网设备）

> 目的地：管道的端点，所有加载数据（即数据湖、数据仓库或数据可视化工具）

> 流：数据从源流向时的变化目的地（即 ETL 或 ELT）

> 处理：将原始数据转化为可用数据的步骤数据（即批处理、流、ML）

> 工作流：顺序和依赖关系数据管道中的进程

连接后，Torch 将监控管道并显示有关每个组件的信息，如图 5 所示。

HK-Acceldata Torch 还包括一个软件开发工具包 (SDK)，开发人员可以使用该工具包将 Torch API 集成到数据管道的开源工作流管理工具中。

使用 workflow 管理工具管理数据管道

比如 Apache Airflow。在 Airflow 中，数据管道工作流以有向无环图 (DAG) 表示，它定义了任务、它们应该运行的顺序以及它们之间的依赖关系。由于程序化集成，Airflow 将结果返回到 Torch 仪表板，使数据工程师能够查看“客户订单”工作流的数据管道流，该工作流在 Databricks 源和 Snowflake 目标之间传输数据，如图 6 所示。

数据工程师可以深入了解“客户订单”工作流程的跨度，如图 7 所示。仪表板显示事件错误和警告，并提供有关事件执行时间和执行方式的时间线。这使数据团队能够跟踪数据管道工作流中特定跨度的可靠性。

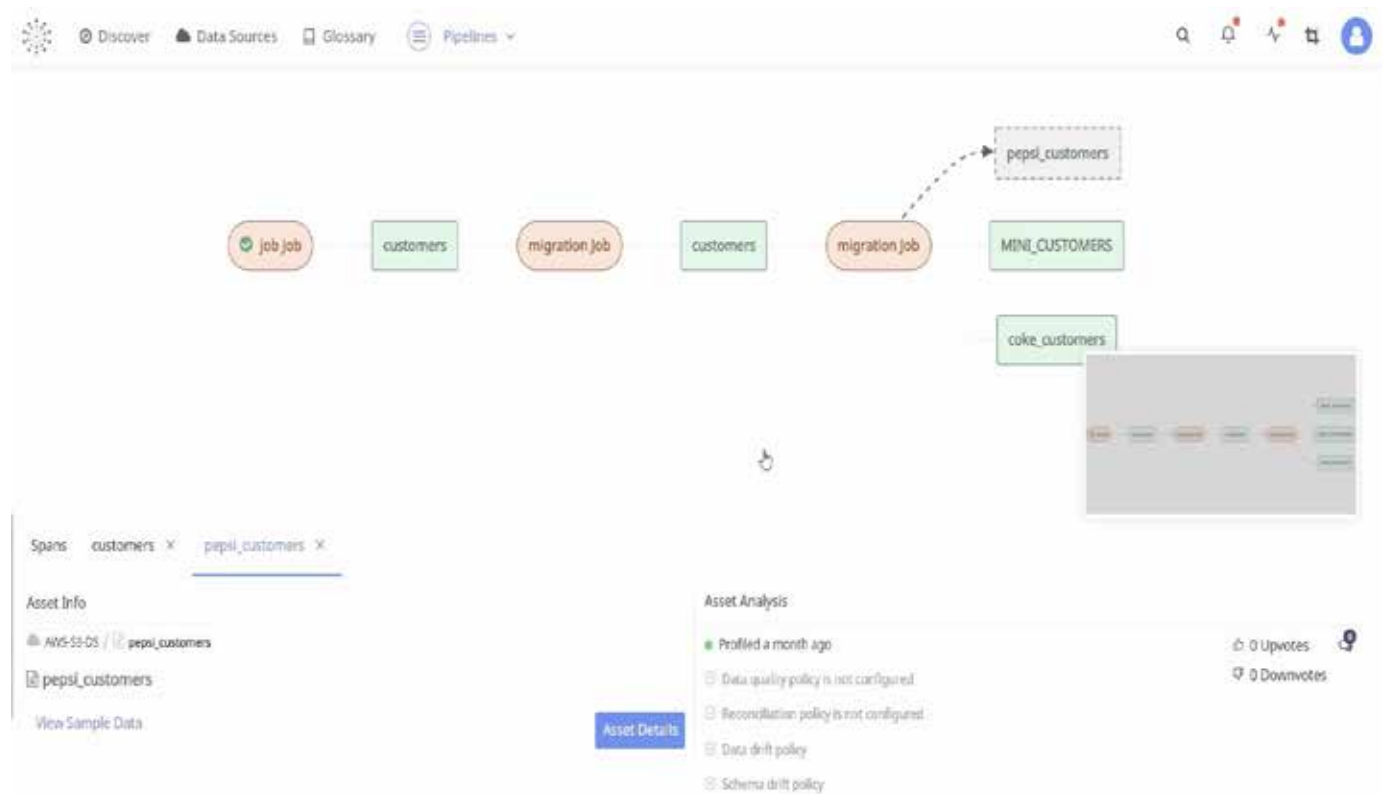


图 6: 数据管道工作流程

跟踪数据沿袭并防止下游问题

HK-Acceldata Torch 跨复杂数据管道跟踪数据沿袭，使数据团队能够可视化和了解数据的来源、数据处理过程中发生的情况以及数据随时间移动的位置。例如，图 8 显示了一个交互式 Lineage 仪表板。

在此示例中，“Customer_Orders”数据资产由三个上游作业（Amazon S3 到 Snowflake 传输作业、Databricks 到 Snowflake 传输作业和 Snowflake 详细信息作业）填充。这些作业正在从更上游的 AWS Redshift 和 Amazon S3 作业接收数据。仪表板还显示“Customer_Orders”数据资产如何影响下游的工作。

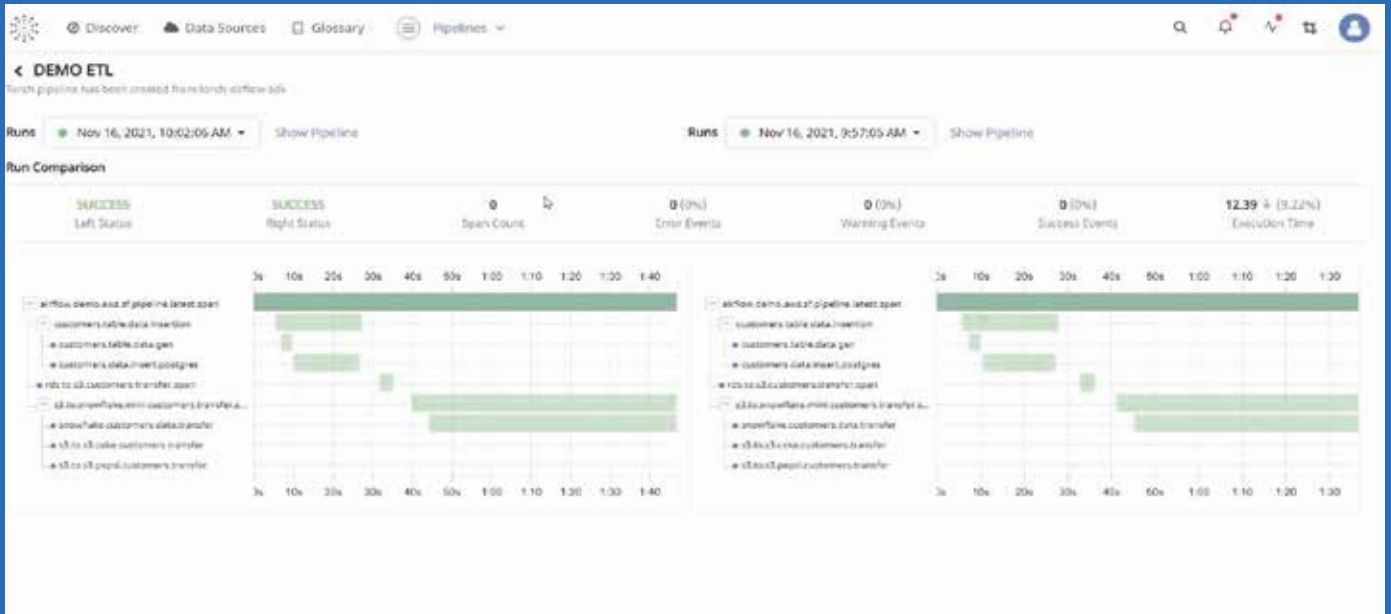


图 7: 查看 workflow 执行范围以发现错误和警告事件

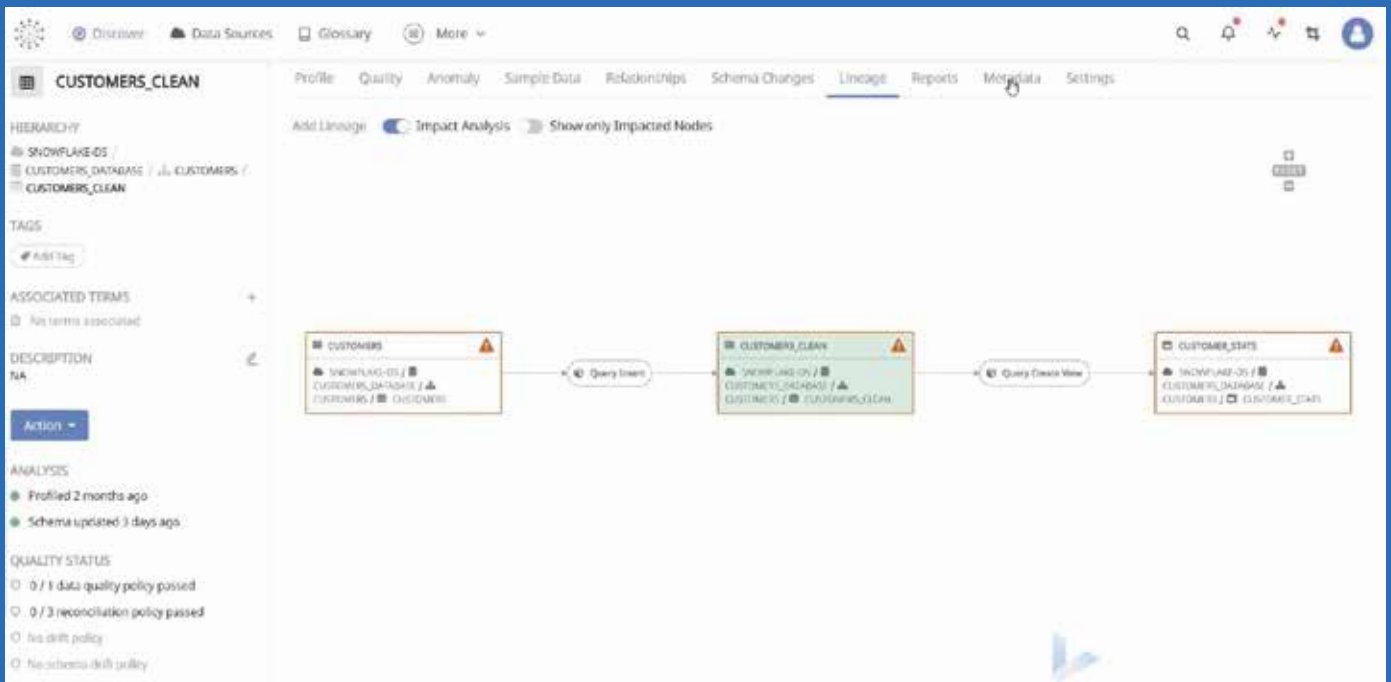


图 8: HK-Acceldata Torch / Asset Details / Data Lineage 窗口

快速解决基础数据或资源问题

HK-Acceldata Torch 与 HK-Acceldata Pulse 预先集成，使数据工程师能够从单一管理平台解决数据管道、数据质量和数据处理问题。通过交互式仪表盘，数据工程师可以深入了解每个处理步骤，并确定哪些事件成功或失败。例如，当数据工程师单击“Customer_Orders”数据集（以绿色表示）时，Torch 会显示架构详细信息，显示数据在流程开始和结束之间是否保持一致。当数据工程师深入研究“Databricks”数据源（以红色表示）时，Torch 会启动 Pulse 以显示与基础架构、容量或内存相关的失败 Spark 进程的信息。借助 HK-Acceldata 数据可观测平台，数据团队可以获得管道可靠性的完整和全面的视图。

简化数据管理

HK-Acceldata Torch 利用人工智能和机器学习来自动化元数据分类、数据分析和数据编目。自动化提高了生产力并减少了错误，同时让用户清楚地了解数据结构、内容和关系。

自动化元数据分类以促进数据使用

HK-Acceldata Torch 使用云和大数据爬虫在首次连接数据源时自动进行元数据分类。数据团队还可以安排定期对数据源进行数据爬取，以监控架构漂移。根据源类型，Torch 收集不同类型的元数据。例如，为数据库和数据仓库收集的元数据可能包括模式、表、列和视图，而用于查询服务的元数据可能包括数据集、视图和查询字段。

自动化数据分析以改善数据访问和质量

HK-Acceldata Torch 执行自动数据分析，以检查和分析数据源中可用的数据。在数据分析期间，Torch 会收集有关数据类型的信息。它收集统计数据，用分类或术语表标记数据，并执行数据质量检查。Torch 在直观的仪表板中显示结果。

集中访问统一数据目录中的数据资产

HK-Acceldata Torch 将元数据和业务术语存储在统一的数据目录中。这有助于集中组织的可用资产库存，无论资产是在云中还是在本地。目录包括：

> 元数据注册表，将数据资产组织在一个层次关系，便于发现。例如：数据库→模式→表→列→字段。

> 帮助用户创建业务的业务词汇表类别并将业务术语链接到数据资产。

HK-Acceldata Torch 通过自动元数据分类、数据剖析和数据编目简化数据管理，以提供组织可用的所有数据的全面视图。

增加数据使用并使数据访问民主化

HK-Acceldata Torch 通过自助发现和丰富工具以及提供丰富可视化的仪表板促进数据民主。

使用搜索和过滤器发现资产和策略

HK-Acceldata Torch 提供搜索和过滤功能，帮助用户查找有关数据资产和策略的信息。用户可以使用以下过滤器执行简单或高级搜索并缩小搜索结果范围：

- > 源类型，例如 MySQL、MongoDB、AWS S3 或 Kafka
- > 数据源连接
- > 资产类型，例如数据库、表、视图、查询或文件
- > #email #company 等标签
- > 诸如“机密性或敏感性：高、中、低”

改善协作和数据发现

HK-Acceldata Torch 使用户能够使用关于资产、所有者或主题的描述来注释数据资产。用户可以添加标签以按用途、所有者或业务功能和评级（赞许或不许）对数据资产进行分类。此外，可以将分类标签添加到元数据中。这种众包信息有助于改善数据发现和整体数据质量。

好处

HK-Acceldata Torch 提供对流经复杂数据管道的数据的质量和可靠性以及数据管道本身的可靠性的可见性。它可以作为独立产品与 HK-Acceldata Pulse 一起使用，或补充现有的数据质量和工作流程解决方案，以提供以下优势：

- > 提供可用数据的 360 度视图的数据类型、数据源或数据位置
- > 通过自动策略检查提高数据质量更好的决策和运营
- > 解决质量、协调或漂移问题，通过自动警报和基于 AI 的建议快速
- > 消除对下游产生负面影响的错误分析和应用

获取演示

获取个性化演示 [请联系我们。](#)

HongKe



虹科电子科技有限公司

www.hongcloudtech.com
hongcloudtech@hkaco.com

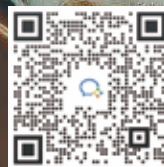
广州市黄埔区神舟路18号润慧科技园C栋6层

T (+86)400-999-3848
M (+86)155 2866 3362

各分部: 广州 | 成都 | 上海 | 苏州 |
西安 | 北京 | 台湾 | 香港 | 美国硅谷



联系我们



行业交流群



获取更多资料



hongcloudtech.com