acceldata

HongKe 虹科

数据可靠性的三个关键支柱

从各种来源收集无穷无尽的数据流，然后将其部署到复杂的管道中，以将这些数据移动到数量不断增加的目标。每一层技术都增加了更多的复杂性，并随之增加了更多潜在的故障点。当数据没有按时到达目的地时，这不仅仅是一种不便——它可能对业务有害。员工、客户、供应商和合作伙伴需要可靠的数据来做出正确的决策，如果没有这些数据，他们可能会采取导致不良结果的行动。

数据可靠性的概念解决了在日益复杂的系统中交付数据的挑战。它建立在多维数据可观测性的范式之上，可以跨复杂环境监控和分析数据操作，以便数据工程师能够识别、预测、预防和解决问题。

“当数据不可靠时，业务主管对不准确的结果感到困惑，数据科学家需要识别替代数据集，数据工程师必须找出瓶颈或损坏数据的来源。”

数据可靠性超越了性能和经典的数据质量问题。数据在流水线中移动时必须进行协调。必须监控可能破坏管道的结构变化（“模式漂移”）。数据还必须跟踪异常和趋势（“数据漂移”），以通知业务或确保分析模型的准确性。

为了确保数据的可靠性，企业必须能够端到端地监控管道，识别数据问题的早期预警信号，快速查明根本原因，并自动化预防性维护以避免业务中断。为了在现代环境中有效运营，企业必须能够平衡数据质量、性能和成本的覆盖范围。洞察数据操作也可用于提高效率。

向实时和连续数据分析的转变

几十年来，业务数据的处理是一个由 IT 驱动的线性过程。熟练的数据工程师构建了相当简单的数据管道，其中提取、转换和加载 (ETL) 以及更改数据捕获 (CDC) 工具从数据库和应用程序中提取原始数据。数据被加载到数据仓库中，并按预先安排的批次进行处理，通常每小时或每晚一次。当单个业务用户或部门请求报告时，数据分析师将使用商业智能 (BI) 软件来执行数据库查询并创建一个静态报告，该报告通常需要几天时间才能构建 - 基于历史数据和过去的结果。如果提出请求的个人或部门对报告有后续问题，它将被放置在请求队列的底部，整个过程将重新开始。

尽管此过程简单明了且易于管理，但它也造成了效率低下的情况。数据通常存放在信息孤岛中，并依靠熟练的数据专家来提取和分析，以便对业务用户有意义。这些限制以及响应新的报告请求或修改所需的时间

数据集市（通常以几周和几个月为单位）导致报告周期缓慢得令人沮丧，这使得实时利用数据变得不可行。

在过去的二十年中，可用数据的爆炸式增长刺激了企业中数据使用方式的巨大转变。由于技术发展的融合，每个行业的企业都从一系列业务活动中收集了无数且不断增长的数据点。当数据管道可以扩展以满足数据及其众多来源的数量和复杂性时，企业可以准确了解其业务运营。凭借这种洞察力，数据工程师可以快速适应以利用他们所学的知识。这些通常是根据增长和市场执行情况显示出超额回报的企业。

如果数据管道可以扩展以服务于数据量和数据的复杂性和许多来源，企业可以根据他们业务运营的准确情况进行操作，并且可以快速适应以利用所学知识。

在这种环境下，企业已转向实时和持续的数据分析，以有效地传递信息并提高业务决策能力。实时数据分析消除了传统商业智能流程特有的数据收集和处理之间的时间延迟。企业现在可以在获得数据后立即访问数据。

随着企业努力将其数据操作化，他们正在将对 BI 和分析的访问扩展到企业的每个角落。不再只是分析师和高管的领域，数据现在可以与一线员工、业务合作伙伴和客户自由共享。业务数据分析集成在整个业务运营范围内，以优化流程、决策、客户关系和产品。

实时和连续数据分析的优势显而易见。能够快速挖掘和分析数据的企业能够做出更快、更好的决策、改进其运营流程、降低安全风险，并最终增加各自市场的收入和数字差异化。

但随着数据量和分析需求的增长，数据管道和处理引擎的复杂性也随之增加。现代管道由各种组件构成，包括提取/加载/转换 (ELT) 和变更数据捕获 (CDC) 工具、在线事务处理 (OLTP) 系统、应用程序编程接口 (API) 和事件流平台，例如 Apache Kafka。他们从应用程序、网站、第三方数据馈送、社交媒体、数据库 IT 日志和物联网 (IoT) 设备中提取结构化、半结构化和非结构化数据。数据被转换并存储在数据仓库、数据湖、NoSQL 数据库和事件流平台中，并通过机器学习算法进行分析。见解，以仪表板和报告的形式，

现代数据管道不是在使用商品硬件的本地环境中运行，而是利用云的可扩展性和成本效益。这要求云对象存储和计算引擎与传统的本地系统集成，进一步增加复杂性。结果通常是一个脆弱的数据管道，必须将数据提供给不断扩展的目标集，包括 BI 工具、人工智能 (AI) 和嵌入式分析工作流。

acceldata

数以千计的交易者管理着$35吨我们平台上的资产。我们无法承受一秒钟的停机时间来从我们的数据中获得洞察力。

数据可靠性定义

数据可靠性旨在确保质量数据的可靠交付、准时处理和大规模端到端管道。

区分数据可靠性和数据质量非常重要。经典数据质量是衡量特定数据集在满足其用户需求方面的适合程度。当满足一系列要求时，数据被认为是高质量的，其中一些要求包括：

(>)**准确性**——数据不含错误，传达真实信息

(>)**完整性**—数据集包括所有信息-实现其目的所需的化

(>)**一致性**— 来自不同来源的数据值是相同

(>)**均匀度**—数据中的所有测量值都是统一的，即全部以千克为单位或全部以磅为单位

(>)**关联**— 数据与其预期兼容用途或目的

高质量的数据对于做出良好的业务决策至关重要。如果数据质量低下或可疑，则企业无法完整准确地了解其企业，并且他们可能会进行不良投资、错失创收机会或损害其运营。

然而，在现代数据管道中，数据是不断运动的。当数据通过管道从源流向目的地时，它会经历几个不同的阶段。集成阶段将多个数据源组合在一起。转换阶段是清理和验证数据的阶段。有简单的处理阶段，数据被汇总、聚合和过滤。最后，还有更复杂类型的使用机器学习的处理阶段，例如预测建模。在这些阶段的任何一个阶段，流程都可能失败或变慢，从而阻止数据到达其预期目的地，并对业务造成潜在风险。因此，高质量的数据并不一定能保证数据的可靠性。

数据可靠性的三大支柱

可靠性要求数据管道的端到端可观测性以及预测、预防和优化数据流的能力，同时考虑到瓶颈检测、资源效率以及最终成本。

数据可靠性围绕三个支柱：

1、管道性能管理

当通过管道的数据流受到损害时，它可能会阻止用户在需要时获得所需的信息，从而导致基于不完整或不正确的信息做出决策。在对业务方面产生负面影响之前识别和解决性能问题，企业需要

能够提供管道宏观视图的数据可靠性工具。当数据在不同的云、技术和应用程序之间移动时，监控数据流对企业来说是一项重大挑战。通过单一窗格查看管道端到端的能力使他们能够了解问题发生的位置、影响的范围以及问题的根源。

计算性能监控对于管理和优化管道性能至关重要。为了确保数据的可靠性，数据架构师和数据工程师必须自动收集和关联数千个管道事件，识别和调查异常情况，并利用他们的学习来预测、预防、排除故障和修复大量问题。

计算性能监控使企业能够：

(>)**预测和预防事件——**计算性能监控提供围绕管道性能趋势和其他活动的分析，这些活动是运营事件的早期预警信号。这使企业能够检测和预测异常，自动化预防性维护，并关联促成事件以加速根本原因分析。

(>)**加速数据消耗——**监控流数据的吞吐量对于减少数据向最终用户的交付时间很重要。计算性能监控允许企业优化查询和算法性能，识别瓶颈和额外开销，并利用定制的指导来改进部署配置、数据分布以及代码和查询执行。

(>)**优化数据操作、容量和数据工程-**计算性能监控通过使 DevOps、平台和站点可靠性工程师能够预测满足 SLA 所需的资源来帮助优化容量规划。他们可以使部署配置和资源与业务保持一致需求、监控和预测共享资源的成本，并通过深入了解数据使用和热点来管理管道数据流。

(>)**与关键数据系统集成——**有了正确的可观测性工具，计算性能监控可以提供对 Databricks、Spark、Kafka、Hadoop 和其他流行的开源发行版、数据仓库、查询引擎和云平台的全面可见性。

2. 数据核对

随着数据通过管道从一个点移动到另一个点，它存在到达不完整或损坏的风险。考虑一个示例场景，其中 100 条记录可能已离开 A 点，但只有 75 条到达 B 点。或者可能所有 100 条记录都到达了目的地，但其中一些记录在从一个平台移动到另一个平台时被损坏。为确保数据可靠性，企业必须能够在所有这些记录从源移动到目标目的地时快速比较和核对它们的实际值。

数据协调依赖于自动评估数据传输的准确性、完整性和一致性的能力。数据可靠性工具通过将源表与目标表进行比较并识别不匹配（例如重复记录、空值或更改的模式）的规则来实现数据协调，以进行警报、审查和协调。这些工具还与数据和目标 BI 工具集成，以跟踪端到端的数据沿袭以及数据何时移动，以简化错误解决。

3. 漂移监测

数据变化可能会影响结果，因此必须监控可能影响数据质量并最终影响业务决策的数据变化。数据容易受到两种主要类型的更改或漂移的影响：架构漂移和数据漂移。

acceldata

*架构漂移*指不同来源引入的结构变化。随着数据使用在整个企业中传播，不同的用户通常会添加、删除或更改结构元素（字段、列等）以更好地适应他们的特定用例。如果不监控架构漂移，这些更改可能会危及下游系统并“破坏”管道。

*数据漂移*描述了机器学习模型中任何会降低模型性能的输入数据的变化。更改可能是由数据质量问题、上游流程更改（例如将传感器更换为使用不同测量单位的新传感器）引起的，或者自然漂移，例如温度随季节变化时。不管是什么导致了变化，数据漂移都会降低预测模型的准确性。这些模型使用历史数据进行训练；只要生产数据与训练数据具有相似的特征，它就应该表现良好。但生产数据与训练数据的偏差越大，模型失去的预测能力就越强。

为了使数据可靠，企业必须建立一个规程来监控架构和数据漂移，并在用户影响管道之前向他们发出警报。

全面的数据可靠性必须能够通过以下方式支持企业数据团队：

消除停机时间

跨数据湖、仓库和其他存储库监控企业数据，以消除影响可靠性的问题。

扩展工作负载

确保关键任务数据和工作负载的可用性。

自动化验证

对静态数据和动态数据进行分类、编目和管理业务规则。

acceldata

用例

PhonePe

PhonePe 是世界上最大的数字支付服务之一，每月处理 4 亿笔现金交易，最高可达每秒 1,100 笔现金交易。为了支持其爆炸性的业务增长，PhonePe 需要快速扩展其基于 Hadoop 的集群，同时添加其他开源数据技术，如 Apache Hbase、HDFS、Kafka、Spark 和 Spark Streaming，以运行其大容量、实时支付和现金转移平台。这种基础设施扩展对系统性能和可靠性造成了巨大压力。

HK-Acceldata 为 PhonePe 的Hbase、Hive 和 Spark数据管道带来了实时可见性。

HK-Acceldata 为 PhonePe 的 Hbase、Hive 和 Spark 数据管道带来了实时可见性，使 PhonePe 的工程师能够从单个应用程序监控其整个数据基础架构。PhonePe 的数据可靠性团队能够使用 HK-Acceldata Pulse 监控其现代数据基础架构性能，以便轻松区分由基础架构问题引起的变化和由季节性或与活动相关的激增驱动的变化。提高了对其性能密集型数据工作负载计算性能的可见性，帮助 PhonePe 显着提高了可靠性。该公司已经能够将其数据基础架构平稳地增长 13 倍，从 70 个节点增加到 1500 多个节点，同时在其 Hadoop 数据湖中保持 99.98% 的可用性。PhonePe 还消除了计划外中断和严重级别 1 问题。

TrueDigital

TrueDigital 一直在努力解决其专有数据软件和存储解决方案的性能不佳问题。该公司每天有超过 50% 的数据未经处理，因为其数据基础设施无法足够快地处理数据量。更不用说系统遭受了严重影响处理能力的可靠性问题。

有了 HK-Acceldata，TrueDigital 的 8 PB 数据湖现在可以在使用 Hortonworks 数据平台（Apache Hadoop、Hive 和 Spark）以及 Apache Ranger、Kafka 和开源 HDP 的 100 多个节点集群上平稳运行。TrueDigital 现在每月处理 5 亿次用户印象，并流式传输近 70,000 条消息/秒，所有这些都不会出现一次意外中断或严重级别 1 问题。

PubMatic

PubMatic 的互联网广告平台帮助世界各地的出版商和应用程序开发人员接触到他们的目标受众。其基于 Hortonworks 数据平台的集群非常庞大——数千个节点处理数百 PB 的数据。但这种规模会导致频繁的性能问题，导致平均解决时间 (MTTR) 非常高。尽管依赖于 HDP 和其他 Apache 开源软件，PubMatic 的基础设施和支持成本也高得惊人。

HK-Acceldata Pulse 帮助 PubMatic 的工程师隔离数据瓶颈、自动提高性能并区分强制性和不必要的数据。这使 Pubmatic 能够将其 HDFS 块占用空间减少 30% 并整合其 Kafka 集群，从而节省成本。仅减少软件许可证的总成本每年就节省了数百万美元。即使在削减基础设施的同时，PubMatic 也能够提高可靠性

和规模的数据管道，使其工程师能够专注于支持其关键任务分析业务的增长。在很大程度上要归功于 HK-Acceldata，PubMatic 现在每天处理超过 2000 亿次广告展示和 1 万亿广告商出价，同时处理 2 PB 的新数据。

市场前景

数据可靠性扩展了企业使用的应用程序性能监控 (APM) 工具的功能，以确保应用程序及其相关基础设施以最高效率运行。APM 工具持续监控应用程序环境并利用机器数据来检测异常、识别趋势、优化资源使用并在性能问题影响最终用户之前对其进行故障排除。使用机器学习，APM 工具将基础设施事件关联起来，以查明减速、阻塞和故障的根本原因，然后帮助管理员修复它们以改进和维护运营工作负载。

虽然这些传统的 APM 工具足以提供对微服务和 Web 应用程序的性能和健康状况的可见性，但它们并非旨在支持复杂的数据系统及其众多用例。如果无法在这些复杂的生态系统中关联和关联事件，数据架构师和工程师很难快速识别和解决问题。

数据可靠性还与其他两个细分市场相交。它与 DataOps 共享功能，后者将 DevOps 和敏捷开发原则应用于数据应用程序和数据管道的开发和优化。DataOps 协调人员、流程和技术以快速向用户交付数据，而数据可靠性带来监控、数据验证和沿袭片段，以帮助确保数据质量和数据交付性能。数据可靠性还涉及 ITOps 和 AIOps 的监控、诊断和修复方面，提供、管理、监控和调整 IT 基础架构资源。

如何提高数据可靠性

(>)**开始观察**——增加数据可靠性的第一步是降低数据系统的复杂性。数据可观测性平台将为您的管道提供全面的可见性，无论架构如何，并提高您对处理 AI 和分析工作负载的所有元素的控制。这种端到端流程的高级视图使您能够识别和深入研究导致延迟、故障和其他数据可靠性障碍的数据和处理问题。

(>)**消除数据停机时间**——监控跨混合数据湖和仓库的数据很重要，以确保高数据质量和可靠性。计算性能监控通过关联整个环境中的事件以进行快速根本原因分析、分析性能趋势以预测潜在故障以及自动修复以防止事件影响操作，从而提高数据处理的可靠性。

(>)**自动化数据验证**——数据漂移会影响人工智能和机器学习的准确性。通过解决数据质量、架构漂移和数据漂移的持续自动化验证，在影响运营之前检测漂移非常重要，以消除中断并提高分析和 AI 的准确性。

(>)**监控动态数据**——数据从来都不是静态的。企业需要使用与数据源、基础架构和云提供商无关的企业架构，对整个数据管道中的动态数据进行分类、编目和管理业务规则。

acceldata

使用 HK-Acceldata 提高数据可靠性

HK-Acceldata 推出了市场上第一个多维数据可观测性云，为采用混合数据湖和云数据仓库的企业提供易于访问、即时可用的服务。HK-Acceldata 多维数据可观测性平台使企业能够利用实时可观测性来：

(>)跨环境（例如混合数据湖和仓库）构建、操作和优化复杂的数据系统以及具有最高数据团队生产力和数据投资回报率的云提供商。

(>)在不损失数据质量的情况下快速扩展技术、工作负载和应用程序。

(>)调整数据和业务战略，以确保企业满足经营目标。

(>)最大限度地利用内部数据团队的专业知识，用更少的资源来做更多的事情

获取演示

获取个性化演示*这里。*