2023年非结构化数据管理报告

Original 常华Andy Andy730 2023-09-20 07:30 Posted on 上海

收录于合集
#数据管理/数据分析

40个

Source: The 2023 Komprise Unstructured Data Management Report, September 2023

- **2022年**：2022年非结构化数据管理现状报告
- **2021年**：非结构化数据管理现状报告（美国和英国）

执行摘要

在短短几个月内，科技领域发生了翻天覆地的变化。AI商业模式和与生成式AI相关的新产品如雨后春笋般涌现。当前的AI浪潮迅速塑造了新的工作方式，带来了显著的生产率提升，改变了产品和服务的创造与分发方式。根据彭博智能分析数据，生成式AI市场的规模预计将在未来10年内从2022年的400亿美元增长到1.3万亿美元。

当然，新型AI也伴随着一系列潜在危险。从隐私和安全风险到伦理考量以及来自不准确或带有偏见数据的风险，政府和企业领袖正在审视这些问题，并权衡采用AI技术在我们社会中的安全和成功实施的解决方案。

在这一巨变中，我们的2023年研究发现，IT领袖将专注于以下三个核心领域：

- 为AI做好准备；
- 为部门用户提供数据服务，如文件搜索和标记；
- 采用云成本优化策略。

这些趋势的融合反映了非结构化数据管理的成熟。IT和存储主管现在被期望与高优先级的企业AI和数据驱动型倡议密切协作。

报告亮点

为AI做好准备

- 2023年，为AI做好准备成为首要的数据存储优先事项，占比高达31%，紧随其后的是云成本优化；

- 绝大多数组织（90%）允许员工使用AI，并且多数（65%）已经制定了相关政策来监管AI的使用；
- 企业AI使用的首要关切是侵犯隐私和安全问题，其次是数据溯源以及因不准确或带有偏见的数据而带来的风险。

从存储管理到数据服务

- 85%的受访者认为非IT用户应该在管理自己的数据中发挥一定作用，而62%已经实现了一定程度的非结构化数据管理用户自助服务；
- 在重要的非结构化数据管理能力中，容量问题和异常监控与警报占据领先地位，占比高达44%；
- 最重要的非结构化数据管理挑战是在不干扰用户和应用程序的情况下迁移数据，占比高达47%，紧随其后的是为AI和云服务做好准备，占比46%。

更多数据，更多支出，新的目标

- 50%的组织正在管理5PB或更多的数据，与2022年相似；
- 管理10PB以上数据的组织比例从27%增长到32%，增长了19%；
- 近三分之二（73%）的组织在IT预算中花费30%或更多用于数据存储和保护，明显高于2022年的67%。

I. 数据存储优先事项：为AI做好准备和云成本优化

2022年，我们报告指出，大多数（65%）组织计划或已经将非结构化数据传递到其大数据分析平台。虽然这些分析平台很可能包含AI，但当今的趋势明显偏向AI。在ChatGPT席卷全球之后，生成式AI立即成为每个董事会、教室和政府办公室都在讨论的技术。而在2023年，当被问及未来12个月内的首要数据存储优先事项时，31%的受访者表示为AI做好准备，其次是云成本优化（22%）。值得注意的是，在2022年，参与者主要关注云迁移，有**56%**表示这是他们的首要举措。其他优先事项包括投资数据管理和移动性，购买更多本地存储设备以及现代化备份和灾难恢复。

未来12个月内的首要数据存储优先事项

- 为AI做好准备：31%
- 云成本优化：22%
- 将更多数据迁移到云中：18%

洞见：转向AI需要成本优化策略

如今，IT团队在不确定的经济环境中面临着减少云和存储支出的压力。在大流行期间，云支出并没有总是带来预期的投资回报，因此在2023年初，人们开始讨论云"遣返"的话题。云成本优化需要采取策略，例如使用第三方工具监控支出、通过自动发现和企业政策管理云扩张以及防止影子IT的出现。数据管理也至关重要，因为它使存储和IT管理人员能够查看所有存储中的数据资产，并将数据放置在当前需求的最佳存储解决方案中。

消除存储浪费对于为蓬勃发展的AI倡议腾出投资和资源至关重要。从数据存储的角度来看，为AI做好准备意味着获取适当的高性能、可扩展的本地和云存储技术组合，同时尽可能具有成本效益。

II. 数据管理对AI的影响

为AI做好准备，不仅仅意味着构建AI-ready的存储基础设施，选择适当的工具同样至关重要。可以说，现在市场上已经有了丰富的价格合理的AI工具和服务，可以满足几乎每个行业的需求。主要的云服务提供商和知名的企业软件供应商都在推出自己的生成式AI解决方案。然而，在部署任何工具或服务之前，IT必须深思熟虑地考虑数据管理对这一过程的影响。

尽管那些多样、易用且引人入胜的生成式AI新场景令人兴奋，但背后也伴随着一系列令人担忧甚至可怕的问题。这些问题包括从敏感数据泄露到威胁公司知识产权和个人身份信息保护的通用语言学习模型（LLMs），再到伦理、准确性、数据源溯源的担忧，以及派生作品的版权问题，甚至是恶意行为者可能造成的威胁，这些都是企业和社会需要认真面对的问题。

洞见：采用生成式AI，但需要为工具和数据设定保障措施

这些担忧正在影响着企业对于采用AI的决策，但根据我们的调查，这并没有阻止商业和IT领袖追求这项技术。调查中，对生成式AI的计划有各种各样的回应，这在早期阶段并不令人意外，然而多数人（44%）限制了员工可以使用的工具和/或数据。

企业AI计划

- 我们尚未制定政策，因为我们仍在努力弄清楚：26%
- 任何员工可以使用任何数据，但仅限于经批准的AI服务：24%
- 任何员工可以使用AI来提高他们的生产力，没有限制：21%
- 只有一些企业数据可以与经批准的AI服务共享：20%
- 我们不允许员工使用AI：10%

生成式AI的数据管理主要关切

IT领导者需要考虑来自生成式AI的一系列隐私、安全、法律和伦理问题。他们最担心的是防止安全和隐私违规（28%），其次是提高数据源透明度以避免生成式AI工具产生不道德、偏见或不准确的输出（22%）。

- 违反我们数据的隐私和安全：28%
- 数据源溯源：缺乏数据源透明度和/或供应商通用语言学习模型（LLM）中不准确或带有偏见的风险：21%
- 从通用LLM衍生作品的法律模糊性：16.6%
- 企业数据泄露到供应商的语言学习模型（LLM）中：16%
- 使用包含他人个人身份信息的GenAI输出可能存在的潜在责任：10.8%
- 我们没有担忧或不确定：7%

减轻AI中非结构化数据的风险需要采用多种策略，但不采取任何行动不是一个选项。

- **40%**表示他们将采取多管齐下的方法，包括存储、数据管理和安全工具；
- **35%**将与其现有的安全/治理供应商合作；
- **32%**表示他们的数据存储和/或非结构化数据管理解决方案具备相关能力；
- **31%**已经创建了一个内部工作组来制定和执行策略，以及；
- **26%**将仅与具备足够保护和控制措施的AI供应商合作。

洞见：关注治理以实现AI的成功

显然，对非结构化数据治理议程的需求很强烈，因为IT领导者不能忽视数据的完整性、数据保护以及生成式AI项目可能出现的故障或危险的结果。Wakefield Research和Informatica的一项2022年调查也证实了这一点，将数据治理列为首席数据官（CDO）的首要任务。

III. 数据管理自助服务日益流行

IT自助服务的整体趋势现在已经扩展到非结构化数据管理领域。企业内的部门IT经理、数据分析师、研究人员和其他数据相关利益相关者都被鼓励积极参与管理自己的数据。这种做法有助于促进更好的合作，让各方一同决定哪些数据应存储在高性能存储设备中，哪些数据可以进行归档或删除，并且允许执行任务如为文件添加元数据和跨数据隔离进行搜索等操作。

目前的趋势显示（36%）是允许员工查看分析数据，例如部门数据增长、文件类型和使用趋势，可以在数据隔离中搜索数据，并创建自定义工作流程。然而，IT希望保持一定的控制：只有**22%**的人认为用户应该能够完全管理他们的数据，包括分层、删除、迁移、召回等操作。

您希望部门用户如何管理他们的数据？

- 用户应该能够查看关于他们的存储使用情况的分析数据，搜索并找到他们需要的数据，并创建自定义工作流程和数据服务：35%
- 用户应该有一定程度的访问分析数据和搜索数据的权限，但IT应该管理其他所有事项：27%
- 用户应该能够自主管理他们的数据，包括分层、删除、迁移、召回、搜索和获取数据分析：22%
- 用户不应该具备任何数据管理权限：16%

洞见：数据服务101：用户追求更多数据控制权

数据服务的范围包括：

- 管理数据在其生命周期内的过程；
- 对数据存储增长和成本进行分析和报告，包括部门级展示和数据使用情况；
- 文件搜索和标记；
- 数据迁移、分层、复制和删除等数据可移动性场景。

尽管仍有超过三分之一（38%）的人正在构建数据服务策略，但另外的**37%**正在积极推进自助数据管理，分享了部门数据使用情况和支出的展示报告。自助数据管理既有益于IT，也有益于部门用户：前者更容易实现成本节约和合规性目标，而不引发冲突，而后者则可以更多地参与决策，以实现业务目标。

您在实现您的目标方面进展如何?

- 我们仍在构建我们的数据服务策略：37.7%
- 我们分享了按部门展示数据使用情况和支出的报告，但数据管理由IT负责：37%
- 我们已经实现了用户自助数据管理：25.3%

IV. 非结构化数据管理成本、挑战、需求：新基础

数据量和成本

在2022年，50%的组织管理了5PB或更多的数据；这一趋势在2023年的调查中仍然持续。然而，尽管在2022年有27%的组织管理了10PB或更多的数据，但今年重型数据拥有者这一部分已经跃升到了惊人的32%。

究竟什么是10PB的数据?

这很难想象，但它相当于110,000部超高清（UHD）电影，或者是美国国会图书馆存储的数据的一半。事实上，三分之一的受访者存储了如此大量的数据，这应该引起人们对企业非结构化数据规模的关注，以及对IT进行管理、保护和存储的相关负担。

与2022年一样，近70%的企业表示他们今年的支出将比去年更多。

在2023年，73%的组织将IT预算的30%以上用于数据存储，较2022年的67%有显著增加。

洞见：成本优化备受关注

尽管数据存储和备份解决方案的选择比以往任何时候都更多，但IT组织仍然在IT预算中花费了大量资金用于存储。在某些方面，这些成本是合理的：保护数据对业务运营、客户成功和整体增长至关重要。然而，过度采购存储容量以避免任何业务中断、云资源的低利用率以及一揽子式的存储策略也存在充足的浪费。了解数据价值、使用趋势、数据优先级和存储/云经济学可以帮助IT做出更好的决策，平衡性能与节省和可持续性目标。

当前管理的数据量

- <500TB：14%
- 500TB到1PB：16%
- 1PB到5PB：20%
- 5PB到10PB：18%
- 10PB到50PB：16%
- 50PB以上：16%

IT预算用于数据存储和保护的比例

- 预算占比<20%：8%
- 预算占比50%以上：16%
- 预算占比30-40%：19%
- 预算占比40-50%：22%
- 预算占比30-40%：35%

非结构化数据管理的挑战和需求

非结构化数据的不受控制的增长导致了能见度的缺失，使得搜索和决策如何以及在何处存储数据变得困难。它还可能导致因影子数据而产生法律和合规风险，存储和备份成本不断上升，以及部门之间的冲突。大多数组织都面临着多个挑战：解决这些挑战或至少将其影响降至最低对于满足用户需求并帮助从数据中生成新价值至关重要。

在2023年，主要挣扎的问题是在不中断地移动数据（这也是2022年的一项主要挑战）和为AI做好准备。前者涉及到一个常见问题，即用户在数据被迁移到新位置或存档以节省费用后无法找到其数据。非结构化数据管理解决方案可以通过透明地移动数据来帮助用户，以便用户可以像以前一样简单地点击文件链接，应用程序也可以继续以相同的方式工作。

为AI做好准备，也是首要的数据存储优先事项，因为它带来了许多未知因素。AI还需要在合适的高性能、可扩展和高效存储（和计算）技术上进行投资。Tirias Research预测，到2028年，生成式AI数据中心服务器基础设施和运营成本将超过760亿美元。在云中，像亚马逊和AWS这样的提供商已经开发了将AI平台与计算和存储能力捆绑在一起的服务，以使公司更容易采用AI。但随着时间的推移，这些服务可能并不会更便宜。

洞见：为AI做好准备需要对数据资产有深刻的了解

无论选择哪种存储解决方案，控制数据存储成本都很困难：本地、边缘或云中都是如此。原因是数据增长没有减缓的迹象。采用分析、分类和分段数据的工具和实践可以导致细致入微的数据管理策略：低优先级的数据存储在存档存储中，直到需要进行活动使用或可以删除为止，而高优先级的活动数据保留在最昂贵的顶级存储中。独立的非结构化数据管理方法可以在许多组织中将年度存储、备份和灾难恢复成本降低70%或更多。

非结构化数据管理的主要挑战

- 在不中断地移动数据和应用程序的情况下移动数据：47%
- 为AI和其他云服务做好准备：46%
- 增长过快，需要展示成本优化：35%
- 部门和用户缺乏对其存储支出和数据使用情况的可见性，这使得对齐困难：34%
- 法律限制要求对不同类型的数据采取不同的处理方式：30%
- 无法看清我们拥有什么以及我们可以进行分层/移动/迁移的内容：23%

非结构化数据管理软件的未来关键功能

非结构化数据管理领域的企业软件部分正在迅速变化。它应运而生，以解决企业非结构化数据的爆炸性增长以及在混合云存储环境中管理数据的复杂性。非结构化数据管理平台应跨足全存储，以便IT专业人员可以迅速获得洞察力，以做出以数据为中心的决策。

在非结构化数据管理方面成熟的IT组织可以通过正确放置数据实现存储和备份的年度节省，节省幅度可达70%或更多。他们还具备向用户提供可行数据服务的能力，如轻松搜索、自动元数据标记和数据移动，以支持各种场景。

- 在2023年，根据44%的受访者的说法，容量问题和异常的监控和警报被确定为未来软件解决方案最重要的能力。
- 基于策略的自动化，例如将数据移动到冷存储或限制删除，是未来软件需求的第二高需求（41%），其后是为业务IT团队和研究人员提供自助访问的熟悉主题。
- 数据保护是2023年非结构化数据管理的首要新场景，与2022年的调查类似。

洞见：非结构化数据管理解决方案必须从一个平台上满足许多关键目标：从跨数据隔离的整体可见性、监控、搜索和文件标记，到自动化策略、透明地访问数据的任何位置以及数据治理。

非结构化数据管理的未来关键功能

- 容量问题和异常的监控和警报：44%
- 基于策略的自动化，例如将数据移动到冷存储、复制、限制：41%
- 业务IT团队和研究人员的自助访问：39%
- 跨数据隔离的数据标记和搜索：33%
- 全球数据索引以实现全面可见性：28%
- 面向AI/ML应用的数据治理：28%

非结构化数据管理解决方案的首要新场景

- 数据保护：60%
- 使用户能够搜索和对非结构化数据运行分析：49%
- 法律保留和合规性：40%
- 查找并删除数据：30%

报告五大要点

一、为AI做好准备是首要任务，同时也是一项重大挑战。

IT领导人一致认为，为企业数据和存储环境为AI实施做好准备是至关重要的，但在生成式AI领域不断变化的情况下，这是一项具有挑战性的任务。组织应考虑如何最好地管理和理解支持这些应用程序的非结构化数据。对非结构化数据的深入了解将在AI解决方案随着监管和标准的演变而变化的情况下，无论如何，都将具有价值。

二、领导者希望AI有监管措施。

AI的担忧范围涵盖了企业数据泄露、伦理和准确性、数据透明度、衍生作品的版权问题以及恶意行为者的操纵，以造成危害。尽管存在这些风险，大多数领导者支持企业使用AI，但希望限制员工可以使用的工具和/或数据。采用涵盖存储、数据管理和安全工具的多管齐下的风险管理方法很受欢迎。

三、数据服务方法逐渐崭露头角。

从存储管理向数据服务的演变自疫情爆发以来一直在酝酿中。数字业务目标和流程占主导地位。IT不仅需要保护数据并使其容易获得，还需要确保用户和部门负责人能够了解自己的数据使用情况，并参与决策，以管理和移动数据。我们的2023年研究表明，大多数组织在发展成熟的自助数据管理架构方面已经走得很远。

四、在云计算谨慎的背景下，成本优化工具和策略不断增长。

由于云计算对一些组织来说未能如期望地节省成本，IT领导人正在深入研究他们的存储和云投资。今年，组织不再如此专注于积极的云迁移，因为云成本优化已成为更重要的优先事项。持续分析数据资产和支出的工具将更好地理解数据资产，然后理想情况下，按照政策自动采取行动，将数据不断地移动到最适用于其场景的最具成本效益的存储中。

五、可见性和洞察力对于管理数据增长至关重要。

在竞争激烈的数据存储市场中，云提供商和传统存储公司不断提供不同用途的新的成本效益解决方案。这一现实让人们惊讶的是，组织正在将IT预算的越来越高的百分比用于数据存储和备份。IT管理人员需要更多关于成本、数据使用情况和特征的洞察和预测分析，以及对数据进行正确放置的自动化方法。这将有助于为AI、大数据和其他数据项目腾出资金。

---【本文完】---

近期受欢迎的文章：
- 内存架构演进：CXL与RDMA的协同发展
- DPFS: 基于DPU的文件系统虚拟化（论文+PPT）
- 高端存储进化：技术和架构的革新
- 6家存储系统公司的客户反馈（最喜欢的/最不喜欢的）
- 加速GPU与存储或内存之间的数据传输

而是异常复杂的系统工程

需要深度洞察

欢迎一起分享思考和见解

常华Andy

上海 徐汇

扫一扫上面的二维码图案，加我为朋友。

Andy730

收录于合集 #数据管理/数据分析 40

上一篇 · Data Mesh与其它数据管理方案对比

Read more

People who liked this content also liked

【一句】CXL SSD 箭在铉上（几篇文章）

Andy730

平台与数据是数字化转型的两个关键要素

许永硕

云计算——云计算核心技术

网络豆云计算学堂