白皮书

规模化地发掘非结构化数据中的洞察力

赞助方：IBM

Amita Potnis

2018 年 10 月

IDC 观点

数据是一项重要的企业资产，数据的价值与数据的最终使用方式息息相关。当今数字世界是由数据驱动的世界。业务决策和战略由分析数据所得到的洞察力驱动。企业的本质正在转变为提供更高的价值，而其中的关键在于打造无缝的底层 IT 基础架构。数据管理系统提供专业的功能，用于数据的摄取、分类、转换、权限管理、检索和资源优化。

大数据存储库对于企业通过分析收集业务开发所需的洞察力至关重要。但是，只有对数据进行适当的分类和标记以支持搜索和查询，企业才能利用数据存储库获得丰富的洞察力。在没有元数据（分类和标记）的情况下，数据集可能无法用于业务开发目的。

企业已经逐渐适应了非结构化数据增长的速度，并且认识到他们基本上不可能从庞大的数据存储库中获取价值。当数据在不同的存储部署位置（传统位置和私有/公有云）中以不同的数据类型孤立保存时，该问题会变得更加严重。市场需要数据管理工具来简化 ETL（提取、转换和加载）流程，以支持三个主要领域：治理，分析和存储优化。IBM Spectrum Discover 是市场上一款全新的数据管理产品，旨在满足上述需求并支持任意企业实现巩固业务路线图的目标。

本白皮书的内容

本 IDC 白皮书评估了当今市场对数据管理解决方案的需求，并着重介绍了 IBM Spectrum Discover。这是一款新产品，旨在管理基于文件和基于对象的非结构化数据；它也是 IBM Spectrum Storage 产品组合的一部分。

形势概要

如今，我们处于 IDC 定义的第三平台上，第三平台基于四大支柱之上：社交、移动、大数据和云。这些技术推动了数据的激增。在过去的 50 年里，标准普尔 500 强企业的平均寿命从 60 年缩短到了 18 岁。为了生存，企业不仅需要实施数字化转型，还需要同时提高适应能力，以便在市场上保持与时俱进，确保利润率。在过去的十年里，数据的增长源自市场上大量涌现的移动设备、传感器、监控摄像头和消费类相机，以及基于 Web 的社交或事务性交互。企业在多个位置生成和存储数据，包括结构化（块）或非结构化（文件和对象）数据，位置包括：核心、边缘、端点以及各种基础架构平台，包括传统的内部/外部环境，私有云和公有云。市场瞬息万变，不断变化的技术在推动企业实施转型。这种情况下，企业如果不采用最新且不断发展的技能集和智能数据管理工具，他们就会落后于竞争对手。

IDC 预测，到 2025 年，全球将生成和复制 163ZB 数据，数据量相当于 2016 年的十倍。这样的数据增长给 IT 部门带来了巨大的压力，IT 部门必须管理和维护基础架构，并始终确保数据的安全性和可用性。此外，IT 部门在满足最新业务需求和基础架构要求时举步维艰。移动设备和社交媒体的使用使得非结构化数据（例如视频、音频片段、电子邮件和图片）激增。

企业需要长期存储这些非结构化数据，以满足法规或业务要求，并分析数据，挖掘新的业务增长机会。这种情况下，企业将以前所未有的方式增长数据集的规模。基于文件和基于对象的存储预测结果显示，到 2021 年，全球将部署 422EB 的存储容量，以支持基于文件和基于对象的存储环境向上和向外扩展，在 2016-2016 年之间，存储容量的复合年均增长率为 30.6%。

表 1 展示了为了在企业内部及外部存储基于文件和基于对象的数据而部署的原始存储容量。鉴于很大一部分存储容量由基于文件和基于对象的存储环境提供，表 2 展示了企业必须管理的非结构化数据量。

表 1

在企业内部和外部存储中部署的原始容量，以支持块、文件和对象环境

	1TB - 100TB	101TB - 1PB	1PB+
内部部署存储	57.6%	27.1%	15.3%
外部部署存储	58.0%	26.0%	16.0%

来源：IDC 基于文件/对象的存储调研结果，2017 年（调查对象 = 450，调查仅限北美地区）

表 2

基于文件和基于对象的存储提供的原始容量百分比 (%)

	基于对象的存储	基于文件的存储
内部部署存储	35.0%	35.5%
外部部署存储	37.3%	33.5%

来源：IDC 基于文件/对象的存储调研结果，2017 年（调查对象 = 450，调查仅限北美地区）

例如，用于机器学习和人工智能分析或基因测序的分析系统可支持多种用例，包括（但不限于）更智能的/抢占式搜索、自动化元数据生成、更智能的文档分类、数据提取，以及自动化内容密集型流程的优化和决策。同时，IDC 研究表明，企业计划将混合/多云存储战略纳入基础架构路线图中，这会增加出现数据孤岛的可能性。云的普及使得市场对强大的数据管理系统的需求增长，以支持更高的业务敏捷性，更快的 ROI 和更高的盈利能力。随着孤岛式存储基础架构中非结构化数据集不断增加，企业必须有能力实现有序的数据管理，这一点很关键。如果没有合适的工具来管理各个孤岛上的数据，企业可能会增加管理开销，同时也会错失宝贵的数据洞察力。借助利用元数据的工具（包括加速型数据分类、标记和索引），企业能够以更快的速度、更大的规模和更高的效率来搜索、查找、检索和分析数据。随着企业继续推进数字化转型，数据集不断增长，他们需要考虑使用数据管理工具来回答三个重要领域中的以下几个（或所有）问题：

治理

- 谁在访问什么数据，访问频率如何?
- 哪些数据是冗余、过时或琐碎的?
- 数据在安全性方面有多敏感?
- 企业内的谁拥有此数据?

存储优化

- 冗余、过时或琐碎的数据存储在何处?
- 企业目前有多少数据副本?
- 哪些数据经常被访问（哪些数据是热数据），哪些不是?
- 哪些数据是任务关键型数据，哪些不是?
- 企业当前拥有多少数据，数据类型有哪些，以及数据存储在何处?

大数据与分析

- 目前哪些应用在使用/消费此数据，以及如何使用数据?
- 数据从哪里/以何种方式生成?谁在以何种方式使用这些数据?
- 每个文件或对象中都有哪些实体、关键字、概念和构面?

元数据和数据管理工具的重要性

元数据可被定义为描述其他数据的数据。元数据通过关联相关标准来帮助您洞悉和控制数据。系统元数据能提供手头数据的更多描述信息，比如创建日期、作者、文件大小、位置和修改日期。系统元数据通常在范围和使用上受限，因为它没有描述数据文件内部的内容。自定义元数据充实了上下文，允许您更具体地整理或组织数据。借助系统元数据，您还可以在更宏观的层面整理或组织数据……除了系统元数据之外，自定义元数据还能简化并增强收集、维护、搜索、集成和分析数据的流程，从而支持您成功实施机器学习/人工智能、生命科学研究等计划。

借助应用和自定义生成的元数据，企业能以简化的方式更快地运行查询，从而支持治理、基础架构优化和分析计划。企业选择通过以下方式，来利用智能数据管理工具在三大领域交付价值：

- **治理**要求企业严格监控和审核谁有权访问哪些数据集并控制数据所有权、角色和职责，以确保数据安全性和合规性。理想的数据管理解决方案将为企业整合与数据治理相关的人员、策略和流程提供最佳解决办法。
- **存储优化**要求企业能够按使用情况和/或策略来识别或标记数据，并将数据移动至适当或指定的存储层，进而以最佳方式使用可用资源。通过利用集中式元数据管理技术，企业能够获得不同存储层（包括基于文件和基于对象的存储）上数据的统一视图。数据管理工具应能够识别和报告：
  - 冗余数据并减少其副本数量
  - 过时的数据并清除这类数据
  - 琐碎的数据并将其清除或将其分层，保存至适当的存储层
- **分析**要求 ELT 流程以闪电般的速度交付，从而分析大型数据集并找到关联性和隐藏模式。

IBM Spectrum Discover

企业正想方设法管理数据集并从中获取价值，为此，供应商通过提供适当的数据管理工具来响应企业的这一需求。IBM 最近宣布推出 IBM Spectrum Discover，该工具可激活数据中的隐藏价值，按数据类别提供存储利用率的可视化视图，并且能够自动捕获和索引系统元数据，同时支持自定义的业务导向型数据标记操作。

IBM 是一家成熟的存储系统厂商并且久负盛名。IBM Spectrum Storage 是 IBM 推出的综合型产品组合，其中包括软件定义存储基础架构、存储服务和数据管理解决方案。IBM Spectrum Discover 通过新的数据管理功能扩展并增强了本就强大的产品组合，这些功能旨在满足企业日益高涨的以下需求：从文件和对象存储中挖掘元数据驱动型数据洞察力。

IBM 认识到市场需要基于元数据的搜索和治理工具后，推出了 IBM Spectrum Discover。IBM 希望该产品能够帮助企业在元数据生成之时就捕获元数据，实现非结构化数据的自动编目，还能帮助企业将自定义元数据标记与系统元数据相结合，以提高可视性，加强数据控制。

IBM 宣布于 2018 年 10 月发布该产品，该产品于 2018 年 11 月全面上市。该产品包括以下功能：

- 基于策略的元数据标记，允许规则识别数据并应用自定义标签
- 能够基于用户的预定义模式，自动和/或手动应用标签
- 元数据标签可以是开放的，也可以是受限的。受限式标签运用预定义值，而开放式标签则运用用户定义的值。
- 基于角色的访问权限，可增强安全性
- 基础和高级搜索功能，包括键值和范围搜索功能，并且可以利用过滤器来优化搜索结果

- 使用 SDK、自定义标签和基于策略的工作流，编排和加速内容识别
- SDK，用于集成第三方应用、开源和/或商用应用
- 直观易用的仪表板和可自定义的报告，支持用户获得详尽的洞察力
- 保存的搜索记录，这样，您无需重新编写查询即可重新运行查询
- 重复的文件检测功能，可用于潜在冗余数据的识别和通知
- 支持您有效使用存储并防止出现数据丢失

目前，IBM Spectrum Discover 能够索引、归类和管理存储在 IBM Spectrum Scale 和 IBM Cloud Object Storage 上的数据。IBM 计划扩展该产品，将公认有竞争力的第三方文件和对象存储产品纳入该产品中。IBM Spectrum Discover 每秒可扫描 30,000 条记录。在创建、更新或删除文件/对象时，该产品还可以利用实时事件通知来更新其数据库。IBM 认为这可以让各行各业受益匪浅，但是他们计划在此初始启动阶段先瞄准医疗保健、金融和电信行业。最终用户将根据所管理的数据量（单位：TB）付费。

IBM Spectrum Discover 的目的是帮助企业改善以下方面：

- **治理。**降低风险，并改善数据质量和生命周期管理。
- **存储优化。**通过对数据进行分层，保存到适当的存储层，并消除冗余数据来提高存储利用率，从而降低成本。
- **大数据与分析。**通过改善和增强分析工作流，深入了解大量非结构化数据，从而更快速地获得准确的结果。

总体而言，该产品旨在以简单、易用的方式解决这些数据管理难题，让最终用户能够采用 IBM Spectrum Discover，并立即从该产品的功能中受益。

挑战/机会

IBM 的优势在于其成熟且多样化的产品组合，不仅包括 IBM Spectrum Storage 产品组合，还包括 IBM 在云和分析领域推出的产品。该公司不仅从系统和软件角度在基础架构领域积累了专业知识，还在服务市场方面积累了专业知识。IBM 意识到市场当前的需求，于是开发了可满足当今和未来需求的解决方案，借此与市场需求保持同步。考虑使用 IBM Spectrum Discover 的企业，可以充分利用 IBM 对数据管理之外的需求的广泛了解和专业知识。

除了现有的功能外，企业还应考虑 IBM Spectrum Discover 将为现有 IBM 产品和平台提供的支持。例如，未来 IBM Spectrum Discover 用户可以利用 IBM Watson Data Platform，搜索和分析来自 IBM Spectrum Scale 和 IBM Cloud Object Storage 的相关数据。

目前，IBM Spectrum Discover 为 IBM Spectrum Scale 和 IBM Cloud Object Storage 数据源提供现成的支持。支持任意存储以确保长期成功，这一点对于所有数据管理工具来说都很重要，IBM Spectrum Discover 提供了一个可扩展的平台来通过 SDK 支持其他数据源。IDC 预测，随着该产品扩展其功能以支持跨越部署位置（传统位置或私有/公有云）的异构存储产品，IBM Spectrum Discover 将迅速深入市场。

结语

维护详细的自定义元数据有助于任何企业实施有关治理、节省成本或分析的计划。通过维护元数据，任何企业都能进一步确保全面实现广泛数据资产的所有业务价值。因此，数据管理工具不仅应纳入数据治理战略，还应纳入长期业务路线图。

企业无法预测非结构化数据的增长速度，因而也无法敏捷地分配基础架构资源。鉴于数据有望继续增长，这还只是一系列数据管理挑战的开始。企业正在设法从数据中获取价值，从而推动企业发展，而为了执行这些活动，企业需要立即使用数据管理工具。我们鼓励用户执行概念验证并评估 IBM Spectrum Discover 可以带来的价值。在一定程度上，IBM 可以应对本文所述的挑战，IDC 认为 IBM 在数据管理市场上有巨大的成功机会。

关于 IDC

International Data Corporation (IDC) 是全球信息技术、电信及消费技术市场领域市场情报、咨询服务与活动的领先提供商。IDC 已帮助许多 IT 专业人士、企业高管及投资社区在技术采购和业务战略的决策方面提供了基于事实的建议。超过 1,100 名 IDC 分析师已在全球 110 多个国家/地区就技术及行业机遇和趋势为其客户提供了全球性、区域性和本地性专业咨询服务。50 年以来，IDC 为客户提供了大量的战略洞察力，帮助客户实现了关键业务目标。IDC 是全球领先的技术媒体、科研和活动公司 IDG 的子公司之一。

全球总部

5 Speen Street

Framingham, MA 01701

USA

508.872.8200

Twitter：@IDC

idc-community.com

www.idc.com