IBM Spectrum Discover

针对内部环境和云环境中的异构文件与对象存储提供统一的元数据管理和洞察力

亮点

- 支持内部环境和云环境中的异构文件与对象存储
- 事件通知和基于策略的工作流
- 精细的存储使用量视图
- 快速、高效地搜索数 PB 的数据
- 快速区分任务关键业务数据
- 自动识别特定类型的敏感信息并对其进行分类
- 基于支持文件中用户可定义关键词的出现情况对元数据进行标记
- 基于内容的数据分类

随着数字转型的推进，非结构化数据的数量也在急剧增加。在此情况下，存储管理员很难跟上非结构化数据的快速增长步伐；组织通常会通过增加存储的方式来解决这一问题，这一点毫不奇怪。

尽管存储量是个挑战，但如果所存储数据的可视性不足，无论是对于大规模非结构化数据的存储管理员还是用户，都意味着更大的挑战。存储管理员通常会发现，单单是系统元数据并无法提供有关存储使用量及数据质量的精细化视图，而这种视图却是高效存储优化的关键所在。此外，对于数据科学家、业务分析师和知识型工作者而言，基本系统级的元数据也不足，导致他们需要花费大量的时间来搜索数据才能完成任务。此外，数据监管人员也无法识别包含机密数据或敏感数据的文件和对象（记录）。

为了克服这些非结构化数据方面的挑战，大型企业已开始转向能够提供一流的数据可视性的元数据管理解决方案。如果组织能够清晰地了解他们的非结构化数据，他们就能优化存储系统，减缓风险，并充分利用非结构化数据的价值，进而获得竞争优势和关键数据洞察力。

改善非结构化数据的经济效益、治理和分析

IBM Spectrum Discover 是一款现代化的元数据管理软件，可针对 EB 规模的非结构化存储提供一流的数据洞察力。IBM Spectrum Discover 可轻松连接至内部环境和云环境中的多个文件与对象存储系统，以快速摄入、合并数十亿个文件和对象的元数据并对其进行索引处理，进而在这些存储源之上提供一个丰富的元数据层。通过这些元数据，数据科学家、存储管理员和数据监管人员可以高效地管理海量的非结构化数据，对其进行分类并从中获得洞察力。这些洞察力可帮助您提高大规模分析的速度、改善存储经济效益并减缓风险，实现竞争优势并加快关键研究的进展。

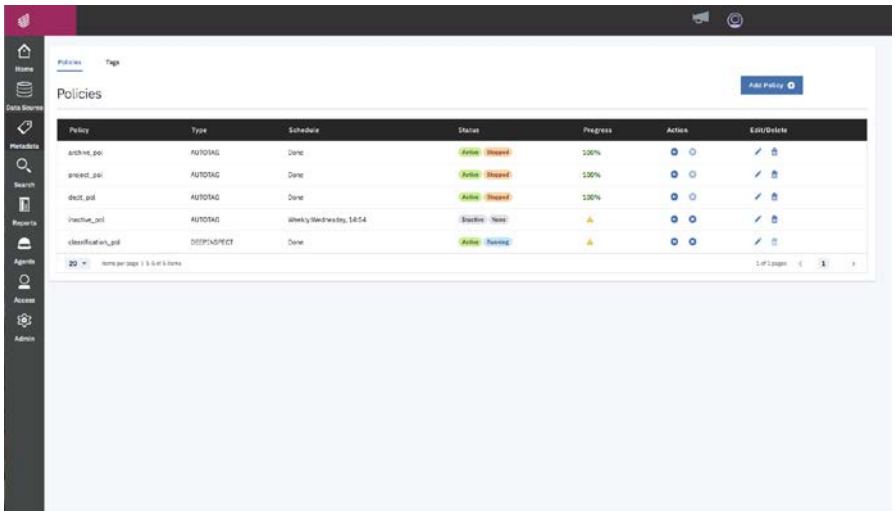
IBM Spectrum Discover 的亮点包括：

- 支持将 IBM 及非 IBM 的存储系统作为数据源，包括 IBM Spectrum Scale、IBM Cloud Object Storage、IBM Spectrum Protect、Dell EMC Isilon、NetApp、Amazon S3、Ceph
- 通过事件通知与基于策略的工作流实现 EB 规模元数据摄入及元数据索引的自动化
- 基于广泛的系统元数据和定制化元数据，提供精细的存储使用量视图
- 可快速、高效地搜索数 EB 的数据，进而生成高度相关的结果用于大规模分析
- 能够快速从拟删除的数据或拟移动至成本更低、使用频次更小的存储层的数据中区分任务关键业务数据
- 基于策略的定制化标记功能可帮助组织对数据进行高效分类，并使其符合业务需求
- 能够基于在受支持文件类型的内容中找到的用户可定义关键词的出现情况对自定义元数据进行标记
- 对特定类型的敏感信息或个人可识别信息进行自动识别和分类
- 通过一个软件开发包 (SDK) 来构建 Action Agent，用于从文件标题和内容中提取元数据，而借助该软件开发包，可实现数据移动的自动化，而且能够实现与 Apache Spark、Apache Tika、PyTorch、Caffe 及 TensorFlow 等开源软件的集成，进而提升数据识别的效率并加快大规模数据处理的速度
- 借助 IBM Spectrum Discover 的应用目录，客户可以发现、安装和管理来自社区支持生态系统的第三方 Action Agent，无需编写自己的代码，就可以扩展 Spectrum Discover 的各项功能

基于策略的元数据标记，可实现精细的数据分类

IBM Spectrum Discover 能够自动从源存储系统捕获系统元数据，通过搜索结果构建定制化元数据，而且允许使用 IBM Spectrum Discover 的 Action Agent API 从文件标题和内容中提取关键词。它能够自动识别可能包含有个人可识别信息（PII）和敏感数据的文档并对其进行自动分类。如此一来，便可形成一个由单个集中式解决方案管理的丰富的文件和对象元数据层。通过对基于内容的数据分类的开箱即用支持，最终用户可以轻松地设置策略，以实现数据的自动识别、分类和目录编制，进而满足特定业务需求。

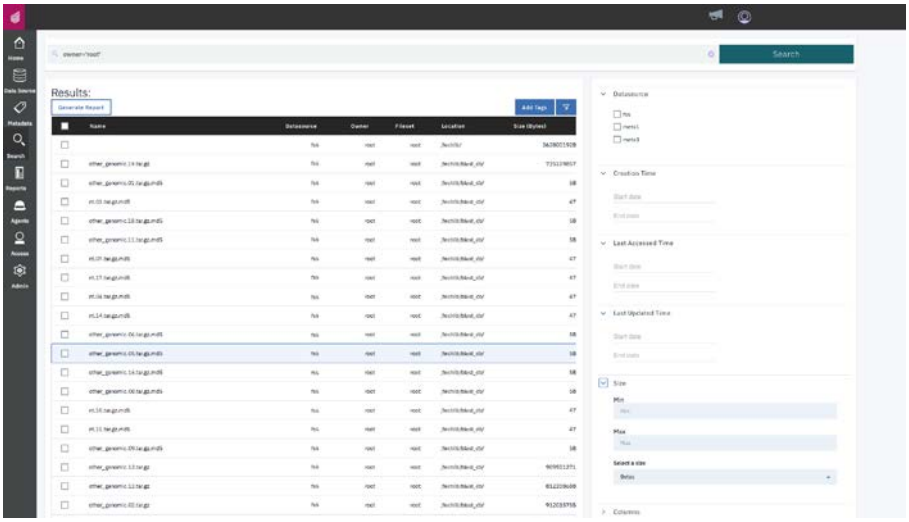
借助 IBM Spectrum Discover，可使用策略实现用于扩充记录元数据的动作的自动化。用户可以将策略运用到任何数据集，而且能够配置动作。举例来说，存储管理员可以轻松地与各个部门协作，对存储时间较长的数据进行归档。若要实现这一点，他们可以使用 Spectrum Discover 的策略引擎，充分利用搜索功能找出某个部门（如营销部门）所拥有的记录，以及在规定期限内（比如超过一年）没有任何访问的记录。之后，他们可以从下拉列表中选择预定义的“归档”标记，而归档标记会自动运用到相关的文件组。策略可以作为一次性事件进行执行，或者预设为定期运行。通过策略创建的任何新标记都会加入索引并变得可以搜索。



IBM Spectrum Discover 提供了一个直观的用户界面，可帮助用户实现基于策略的元数据标记。

通过数十亿个元数据标记实现快速搜索，可帮助您快速发现数据资产

IBM Spectrum Discover 提供了一个搜索栏和一个更高级的搜索面板，可帮助用户快速查找已经过索引处理的记录子集。搜索结果会显示在表格中，而该表格中包含有与搜索标准相关的信息。用户可以查看或不可查看的内容可使用基于角色的访问控制进行决定。

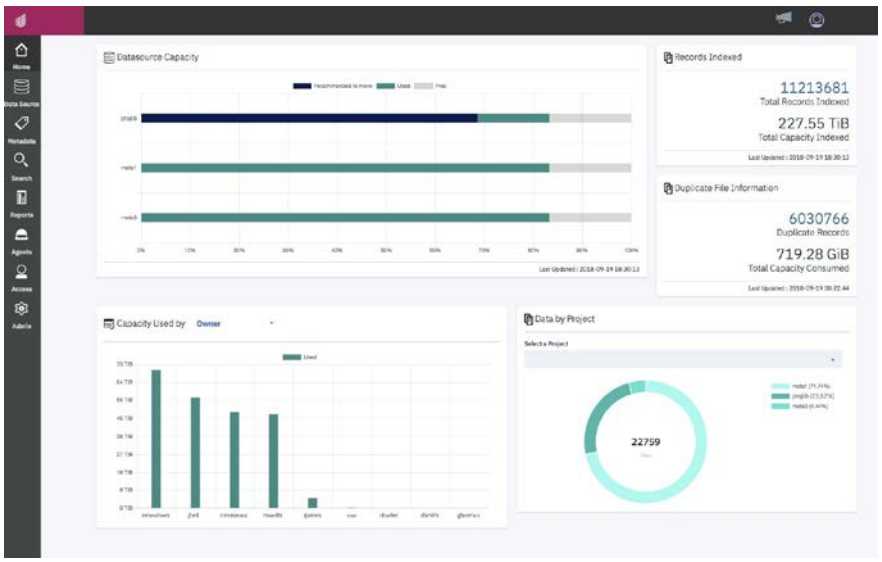


IBM Spectrum Discover 的搜索结果会显示在表格中，而该表格中包含有与搜索标准相关的信息。

熟悉 SQL 语法的用户可以在搜索栏中输入搜索字符串。此外，IBM Spectrum Discover 还提供了一个易于使用的搜索面板，可使用预定义的选择框对记录进行筛选。举例来说，通过“File System（文件系统）”选择框，用户可以选择一个或多个源存储系统。通过“Time（时间）”选择框，用户可以根据记录的上次访问时间规定一个时间范围。通过“Size（大小）”选择框，用户可以根据最小和/或最大文件大小识别记录。借助这些功能，用户可以根据自己的需要采用任意组合的搜索框。

通过仪表板和可定制报告实现记录可视化

IBM Spectrum Discover 仪表板可提供用户环境概览信息。用户可以查看或不可查看的内容可使用基于角色的访问控制进行决定。仪表板中包含有多个小工具，它们会以图形化的方式呈现已由 IBM Spectrum Discover 进行了索引处理的记录的相关信息，使得用户能够实现其数据环境的可视化。举例来说，该仪表板能够显示已注册存储系统的使用情况和容量、潜在重复文件的相关信息，以及按项目或部门统计的容量使用细分信息。



IBM Spectrum Discover 仪表板包含有可显示用户环境概览的小工具。

对于希望获得额外记录详情的用户而言，IBM Spectrum Discover 可为其提供可定制的报告。无论是汇总报告还是详细报告均可生成。汇总报告会按照不同的标准（如对象保险库、文件系统或用户等）汇总记录计数或记录容量等信息并对其进行分组。详细报告会针对系统中符合报告筛选标准的每条记录提供详细信息。

技术规格

单节点试用版	
内存	128 GB（最低 64 GB）
CPU	24 个逻辑处理器（最少 8 个）
存储	
单节点试用版	
基础操作系统和软件	厚置备延迟置零 HDD 或 SSD/闪存 VMDK (500 GB)
持续消息队列	厚置备延迟置零 HDD 或 SSD/闪存 VMDK（50 GB，每 2000 万个已索引文件配备 2 GB）
数据库（包括备份）	厚置备延迟置零 SSD/闪存 VMDK（最低 100 GB，每 200 万个已索引文件配备 2 GB）
数据库（不包括备份）	厚置备延迟置零 SSD/闪存 VMDK（最低 100 GB，每 200 万个已索引文件配备 1 GB）
网络	单 GB 以太网或 10 GB 以太网
单节点生产版	
内存	128 GB
CPU	24 个逻辑处理器
存储	
基础操作系统和软件	厚置备延迟置零 SSD/闪存 VMDK (100 GB)
持续消息队列	厚置备延迟置零 SSD/闪存 VMDK (700 GB)
数据库（包括备份）	厚置备延迟置零 SSD/闪存 VMDK (2.5 TB)
网络	单 GB 以太网或 10 GB 以太网
多节点（3 节点）生产版	
内存	256 GB
CPU	32 个逻辑处理器
持续消息队列	厚置备延迟置零 SSD/闪存 VMDK（每个节点 1.4 TB）
数据库（包括备份）	厚置备置零 SSD/闪存 VMDK（14 TB SAN 存储）
网络	10 Gb 以太网
软件要求	
VMware ESXi 6.0 或更高版本	
受支持的数据源	
IBM Spectrum Scale	
IBM Cloud Object Storage	
IBM Spectrum Protect	
Dell EMC Isilon (NFS)	
NetApp (NFS)	
Amazon S3 (S3)	
Ceph (S3)	

IBM Spectrum Discover 的功能

持续元数据摄入	• 内置连接器可提供与 IBM Cloud Object Storage、IBM Spectrum Scale、Dell® EMC Isilon、NetApp®、Amazon S3、Ceph 的集成 • 事件通知功能可实现持续元数据摄入的自动化（仅限于 IBM Spectrum Scale 数据源） • 元数据索引功能可实现快速数据查询
系统化的元数据监管	• 策略驱动型工作流可实现自动化定制标签功能 • 定制化数据标签可帮助用户精确找出大规模分析所需数据 • 能够将系统和定制化数据标签链接到一起，加速实现存储优化
实时数据洞察力	• 快速搜索功能可在数秒内完成高度相关文件和对象的定位 • 带有向下钻取图标元素的仪表板有助于简化存储管理 • 可定制报告功能有助于加快审计和通信速度
安全且可扩展的架构	• 基于角色的访问控制可确保仅允许经授权的数据访问 • Action Agent API 支持与客户开发的软件和/或第三方软件的集成 • 策略引擎挂钩有助于实现自动化工作流 • 基于内容的数据分类

为什么选择 IBM?

作为业内领先的数据存储产品提供商，IBM 在数据管理解决方案进行了大量投资，旨在改善存储经济效益、数据质量、数据治理水平及数据识别效率，进而为大规模分析和 AI 提供支持。IBM Spectrum Discover 是 IBM 在数据管理方面所具整体优势中的关键一环，可为客户提供强大的元数据管理功能，帮助他们实现数据可视性与分类，进而改善存储优化、提升数据科学水平。

有关更多信息

如欲了解有关 IBM Spectrum Discover 的更多信息，请联系您的 IBM 代表或 IBM 业务合作伙伴，或访问以下网站：ibm.com/us-en/marketplace/spectrum-discover