

IBM Spectrum Discover

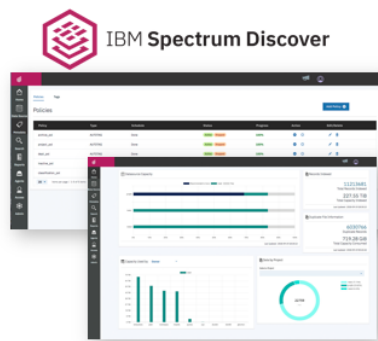
Unlock the value of data and create new insights and real-time analysis

Highlights

- Create custom reports or use interactive GUI
- Locate data in seconds on-prem and in the cloud
- Gain new insight into storage consumption and data quality
- Empower users with self service analysis
- Quickly differentiate, locate and analyze mission-critical business data
- Easily organize, identify and classify sensitive information
- Apply tags based on occurrence of user-definable keywords
- Automatically classify data based on content or metadata tags
- Supports heterogeneous file and object storage on-premises and in the cloud

With digital transformation comes rapid growth of unstructured data. It is no surprise that storage administrators are struggling to keep pace with rapid growth of unstructured data—and that in response, they often find themselves simply throwing more storage at the problem.

Real time data ingest with data mapping, visualization and automatic actions for AI workflows



- **Automate cataloging/indexing** of data by capturing metadata as it's created real-time
- **Locate and identify** the most relevant data regardless of its type or location including CONTENT search
- **Create custom reports** or use interactive GUI interface
- **Enable comprehensive insight** by combining system metadata with custom tags and even content tags based on data patterns
- **Create custom action agents** to automate data workflows

Sample use cases:

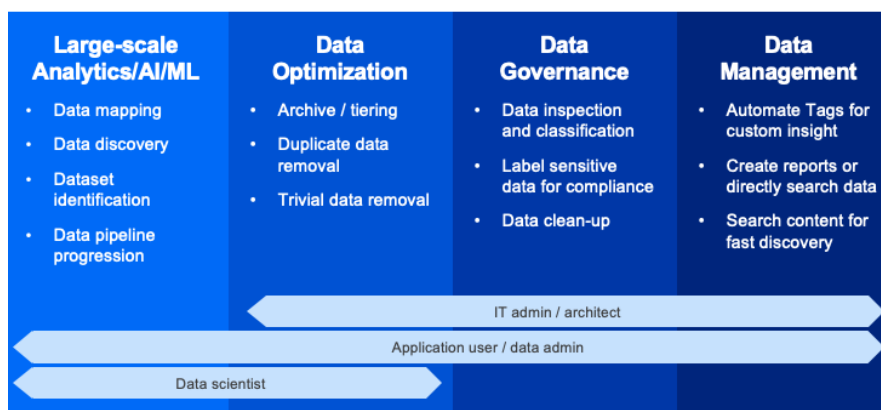
- Faster AI analysis
- Compliance classification
- Image/video indexing
- Identify personal data
- AI data pipeline integration
- Real-time data discovery
- New insights to optimize data
- Find bad or duplicate data

[Spectrum Discover summary](#)

While storage volume is a challenge, limited visibility into stored data poses an even greater challenge for both storage administrators and users of large-scale unstructured data. Users often find that system metadata alone doesn't provide the fine-grained view of storage consumption and data quality that is needed for effective storage optimization. Basic system-level metadata is also inadequate for data scientists, business analysts and knowledge workers who spend a significant amount of their time searching for data necessary to do their work. Data stewards also struggle to identify files and objects (records) that contain confidential or sensitive data.

To overcome these data challenges, large enterprises are turning to metadata management solutions that offer exceptional data visibility. Once organizations have a clear understanding of their unstructured data, they can optimize storage systems, mitigate risk and harness the value of unstructured data for competitive advantage and critical data insights.

Multiple concurrent ways to leverage Spectrum Discover



Multiple ways to leverage Spectrum Discover

There are four main ways to leverage Spectrum Discover

The first way is for large-scale analytics/artificial intelligence (AI) / Machine learning (ML). This includes data mapping, data discovery, data set identification and data pipeline progression. Overall, this approach leverages Spectrum Discover to find out what is in the data.

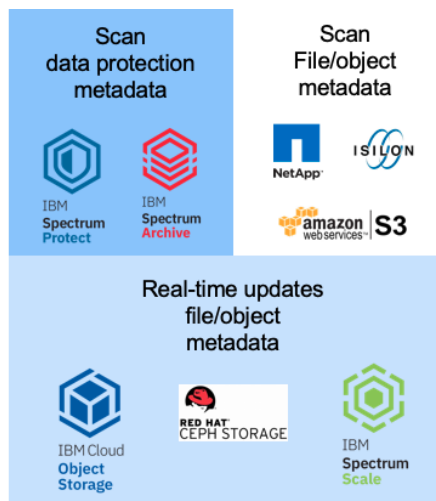
The second approach is data optimization. This includes organizing the data to effectively archive or tiering the data based on the usage frequency of the data. Frequently accessed (i.e. hot/warm) data stays in faster storage devices, while infrequently accessed (i.e. cold/frozen) data are moved into lower-performance storage tiers. Other tasks in data optimization may

involve de-duplication of data and trivial data removal, ultimately reducing the size of the data set.

The third approach is Data Governance. This includes data inspection and classification, labeling sensitive data for compliance and data clean-up.

The fourth approach is Data Management. This includes automatically tagging data for custom insight, creating reports or directly search data, and search content for fast discovery.

Ingest from one or more
unique data sources



Ingest from multiple data sources and types

Spectrum Discover can ingest from multiple types of unstructured data sources. Spectrum Scale, IBM Cloud Object Storage (COS) and Red Hat Ceph 4.0 can all ingest data in real-time, which means that a data scan only needs to occur initially and all subsequent updates to data are automatically sent to Spectrum Discover.

Spectrum Protect and Spectrum Archive are two IBM backup/archive applications recently enabled to integrate with Spectrum Discover. Spectrum Discover can now easily analyze backed up / archived data from those 2 products.

Finally, Netapp, Dell EMC Isilon, and Amazon Simple Cloud Object Storage (S3) can all be scanned by Spectrum Discover to provide a multi-cloud multi-storage metadata repository.

Improve unstructured data economics, governance and analytics

IBM Spectrum Discover is modern metadata management software that provides data insight for exabyte-scale unstructured storage. IBM Spectrum Discover easily connects to multiple file and object storage systems both on-premises and in the cloud to rapidly ingest, consolidate and index metadata for billions of files and objects, providing a rich metadata layer on top of these storage sources. This metadata enables data scientists, storage administrators, and data stewards to efficiently manage, classify and gain insights from massive amounts of unstructured data. The insights gained accelerate large-scale analytics, improve storage economics, and help with risk mitigation to create competitive advantage and speed critical research.

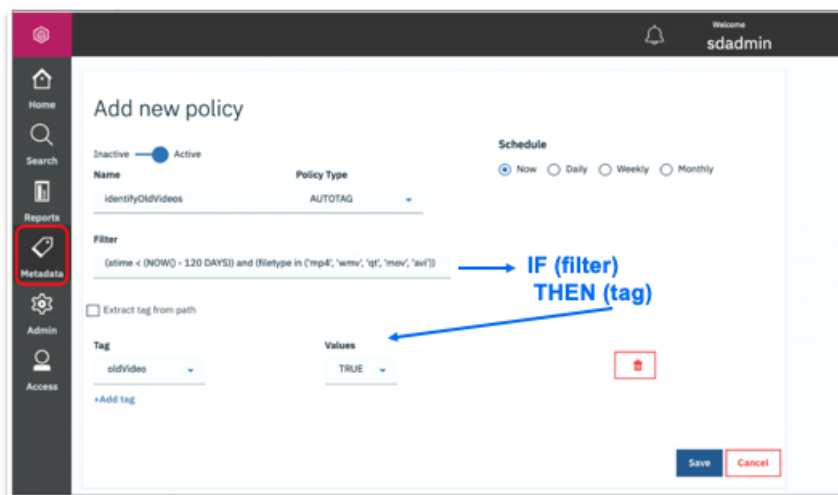
IBM Spectrum Discover highlights include:

- Support for both IBM and non-IBM storage systems as data sources, including IBM Spectrum Scale, IBM Cloud Object Storage, IBM Spectrum Protect, Dell EMC Isilon, NetApp, Amazon S3, Ceph
- Event notifications and policy-based workflows to automate metadata ingestion and metadata indexing at exabyte-scale
- Fine-grained views of storage consumption based on a wide range of system and custom metadata
- Fast, efficient search through exabytes of data, resulting in highly relevant results for large-scale analytics
- Ability to quickly differentiate mission-critical business data from data that can either be deleted or moved to a cheaper, colder tier
- Policy-based custom tagging that enables organizations to classify and categorize data and align this data with the needs of the business
- Ability to apply custom metadata tags based on the occurrence of user-definable keywords
- Automatic identification and classification of sensitive or personally identifiable information
- A Software Developers Kit (SDK) to build Action Agents that extract metadata from file headers and content, automate data movement and provide integration to open source software, such as Apache Spark, Apache Tika, PyTorch, Caffe and TensorFlow, which facilitates data identification and speeds large-scale data processing
- IBM Spectrum Discover Application Catalog enables clients to discover, install and manage third-party Action Agents from a community-supported ecosystem to extend the capabilities of Spectrum Discover without having to write their own code

Policy-based metadata tagging for granular data classification

IBM Spectrum Discover automatically captures system metadata from source storage systems, creates custom metadata from search results and enables extraction of keyword metadata from file headers and content using the IBM Spectrum Discover Action Agent API. Automate the identification and classification of documents that could potentially contain Personally Identifiable Information (PII) and sensitive data. The result is a rich layer of file and object metadata that is managed using one centralized solution. Out-of-the-box support for content-based data classification enables end users to easily set up policies to automatically identify, classify and categorize data, which could be leveraged for specific business needs.

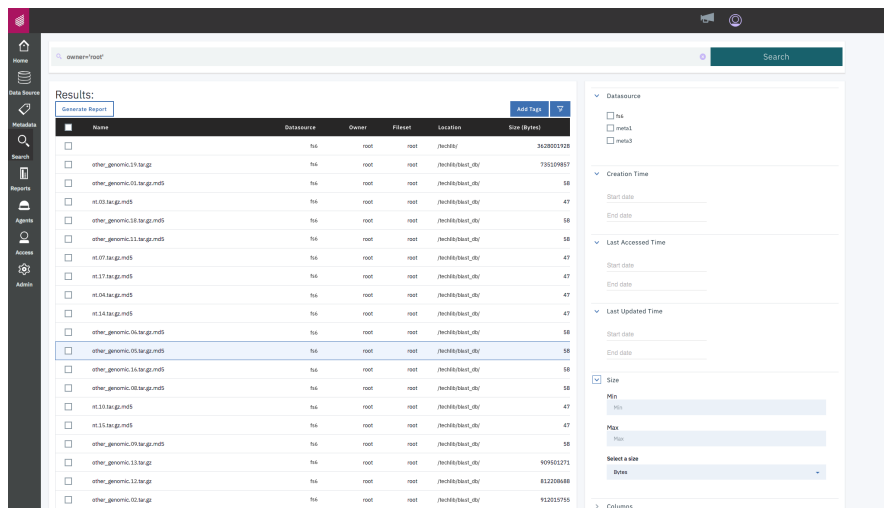
With IBM Spectrum Discover, policies are used to automate actions that enrich metadata. Users can apply policies to any set of records and can configure actions. For example, storage administrators can easily coordinate with departments to archive aging data. To do this, they use the Spectrum Discover Policy Engine that leverages the search function to find records owned by a department (for example, marketing) and that have not been accessed for a specified period of time (for example, more than one year). Then, they select a predefined “archive” tag from a drop-down list and the archive tag is automatically applied to the relevant group of files. Policies can be executed as one-time events, or they can be scheduled to run periodically.



Map metadata with easy GUI interface

Fast searching through billions of metadata tags enables rapid discovery of data assets

IBM Spectrum Discover provides both a search bar and a more advanced search pane to help users quickly find subsets of records that have been indexed. Search results are displayed in a columnar table that contains information correlated to search criteria. What a user can see or not see is determined using role-based access controls.

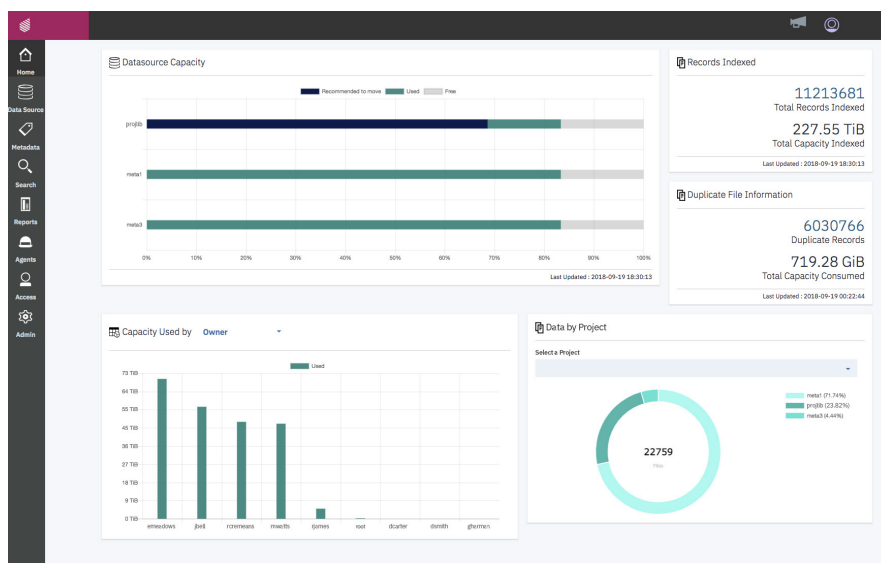


IBM Spectrum Discover search results are displayed in a columnar table that contains information correlated to search criteria.

Users familiar with SQL syntax can enter a search string in the search bar. Or, IBM Spectrum Discover provides an easy-to-use search pane to filter records using predefined selection boxes. For example, the “File System” selection box allows users to select one or more source storage systems. The “Time” selector allows users to specify a range of time based on when records were last accessed. The “Size” selector allows users to identify records based on minimum and/or maximum file sizes. These and other search capabilities allow users to employ any combination of search boxes that best suits their needs.

Dashboard and customizable reporting for record visualization

The IBM Spectrum Discover dashboard represents a user's environment at a glance. What a user can see or not see is determined using role-based access controls. The dashboard contains widgets that graphically present information about records indexed by Spectrum Discover allowing users to visualize their data environment. For example, the dashboard can show usage vs. capacity of their registered storage systems, information about potential duplicate files, and breakdowns of how capacity is being used by projects or departments.



IBM Spectrum Discover Dashboard includes widgets that show a user's environment at a glance.

For users who want additional record detail, IBM Spectrum Discover provides customizable reports. Both summary and detailed reports can be generated. Summary reports aggregate and group information, such as record count or record capacity by different criteria, for example: object vault, file system or user. Detailed reports provide detailed information for each record in the system that matches a report's filtering criteria.

Technical Specifications

Single node trial	
Memory	128 GB (64 GB minimum)
CPU	24 logical processors (8 minimum)
Storage	
Single node trial	
Base OS and Software	Thick-provision and lazy-zero HDD or SSD/flash VMDK (500 GB)
Persistent message queue	Thick-provision and lazy-zero HDD or SSD/flash VMDK (50 GB, 2 GB per 20 million indexed files)
Database (includes backup)	Thick-provision and lazy-zero SSD/flash VMDK (100 GB minimum, 2 GB per 2 million indexed files)
Database (does not backup)	Thick-provision and lazy-zero SSD/flash VMDK (100 GB minimum, 1 GB per 2 million indexed files)
Network	Single gigabit Ethernet or 10 Gb Ethernet
Single node production	
Memory	128 GB
CPU	24 logical processors
Storage	
Base OS and Software	Thick-provision and lazy-zero SSD/flash VMDK (500 GB)
Persistent message queue	Thick-provision and lazy-zero SSD/flash VMDK (700 GB)
Database (includes backup)	Thick-provision and lazy-zero SSD/flash VMDK (2.5 TB)
Network	Single gigabit Ethernet or 10 Gb Ethernet
Multi-node (3 nodes) production	
Memory	256 GB
CPU	32 logical processors
Persistent message queue	Thick-provision and lazy-zero SSD/flash VMDK (1.4 TB per node)
Database (includes backup)	Thick-provision and eager-zero SSD/flash VMDK (14 TB SAN storage)
Network	10 Gb Ethernet
Software prerequisites	
VMware ESXi 6.0 or higher	
Supported data sources	
IBM Spectrum Scale	
IBM Cloud Object Storage (S3)	
IBM Spectrum Protect	
Dell EMC Isilon (NFS)	
NetApp (NFS)	
Amazon S3 (S3)	

Ceph (S3)

IBM Spectrum Discover Capabilities

Continuous metadata ingestion	<ul style="list-style-type: none">• Built-in connectors provide integration with IBM Cloud Object Storage, IBM Spectrum Scale, Dell® EMC Isilon, NetApp®, Amazon S3, Ceph• Event notifications automate continuous metadata ingestion (IBM Spectrum Scale only)• Metadata indexing enables rapid data queries
Systematic metadata curation	<ul style="list-style-type: none">• Policy-driven workflows automate custom labeling• Custom data labels help pinpoint data for large-scale analytics• Ability to link system and custom data labels accelerates storage optimization
Real-time data insight	<ul style="list-style-type: none">• Fast search locates highly relevant files and objects in seconds• Dashboards with drill-down chart elements simplify storage management• Customizable reports expedite audits and communication
Secure and extensible architecture	<ul style="list-style-type: none">• Role-based access control ensures only authorized access to data• Action Agent API supports integration with customer-developed and/or third-party software• Policy engine hooks enable automated workflows• Content-based data classification



Why IBM?

As an industry-leading provider of data storage products, IBM is investing in data management solutions that improve storage economics, data quality, data governance and data identification for large-scale analytics and AI. IBM Spectrum Discover is a key aspect of the overall IBM data management advantage, and provides powerful metadata management that brings visibility and classification to improve storage optimization and increase data science.

For more information

To learn more about IBM Spectrum Discover, please contact your IBM representative or IBM Business Partner, or visit: ibm.com/us-en/marketplace/spectrum-discover

© Copyright IBM Corporation 2020.

IBM, the IBM logo, and ibm.com are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. A current list of IBM trademarks is available on the Web at <https://www.ibm.com/legal/us/en/copytrade.shtml>, and select third party trademarks that might be referenced in this document is available at https://www.ibm.com/legal/us/en/copytrade.shtml#section_4.

This document contains information pertaining to the following IBM products which are trademarks and/or registered trademarks of IBM Corporation: IBM®, IBM Spectrum Scale™, IBM Cloud Object Storage System™.



All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.