Washington Systems Center - Storage

IBM Spectrum Discover

A Technical Overview and Introductory Demonstration of IBM Spectrum Discover 2.0.1

Isom Crawford, *IBM Washington Systems Center*
Norman Bogard, *IBM Washington Systems Center*

The power of data provides significant advantage

Those who harness the power of their data have a significant *competitive advantage* to:

- Predict and shape future outcomes
- Optimize people to do higher value work
- Automate decisions, processes, & experiences
- Reimagine new business models

AI unlocks that value of data to transform business in totally new ways

Source: Worldwide Storage for Cognitive/AI Workloads Forecast, 2018–2022

By 2019…

40%
of digital transformation initiatives use AI

$4.79 billion
AI-related IT storage spend

By 2021…

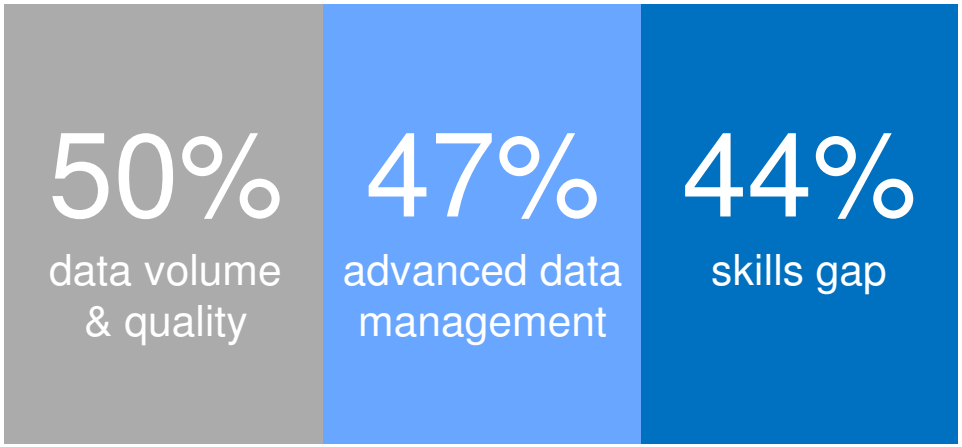
75%
of commercial enterprise apps will use AI

>90%
of consumers will interact with customer support bots

>50%
of new industrial robots will leverage AI

Top 3 Challenges

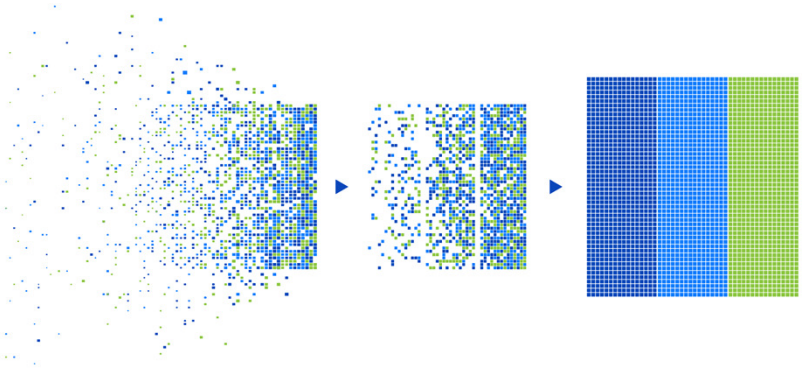
for organizations deploying AI workloads



Source: Cognitive, ML, & AI Workloads Infrastructure Market Survey, IDC, January 2018; n=205, 1,000+ employees (U.S.); 500+ employees (Canada)

Metadata is the key

It brings structure to unstructured data…



… and enables improved data management based on the value / importance of data.

Metadata is data about data:

- Context for data classification & management

Types of metadata

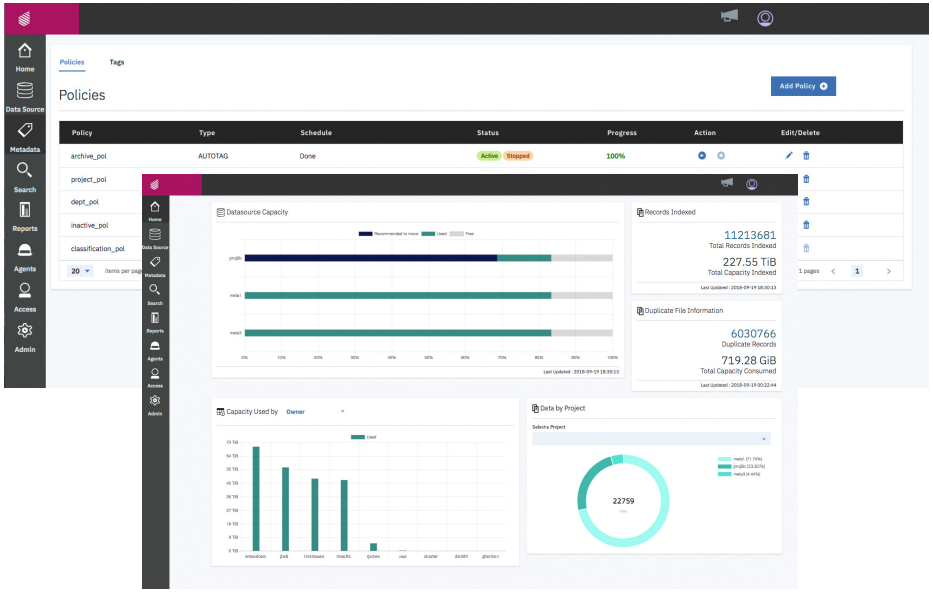
- System: information about file and object types, their sizes, when they were last modified, etc.
- Custom: user/organization-defined based on unique taxonomy
- Derived: derived from analytics and applied to your data enriching the metadata model with additional meaning

Benefits of metadata:

- Identify & manage assets that add value
- Simplify search & access to critical data
- Define & execute policies based on metadata
- Improve time-to-value & storage economics
- Enrich and increases the value of data

IBM’s metadata management solution is the answer

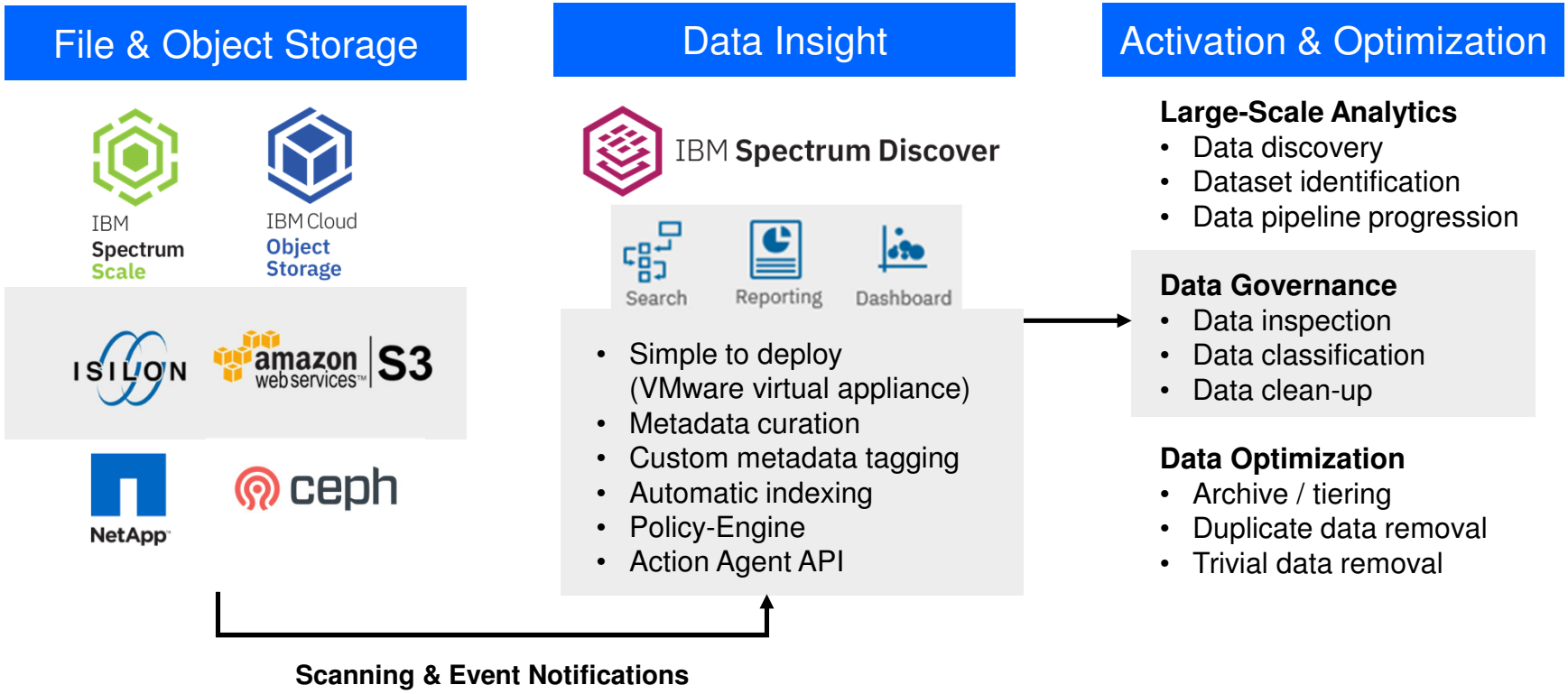
IBM **Spectrum Discover**



Data Insight for Analytics, Governance, & Optimization

- **Automate cataloging** of unstructured data by capturing metadata as it is created
- **Enable comprehensive insight** by combining system metadata with custom tags to increase storage admin & data consumer productivity
- **Leverage extensibility** using the API, custom tags, and policy-based workflows to orchestrate content inspection & activate data in AI, ML, & analytics workflows

IBM Spectrum Discover Overview



Key Features

Connectors	• Scanners for IBM Cloud Object Storage (COS), IBM Spectrum Scale, Dell EMC Isilon (NFS), NetApp (NFS), Amazon S3, Ceph (S3) • Live event notifications for IBM Cloud Object Storage • Tech preview of Notifications (Live Events) for IBM Spectrum Scale
Platform	• Support for Single and Multi-Node (3-Node HA Cluster) (x86 only) • Code upgrade for Single and Multi-node • Encryption of Metadata Database and Notification Logs (Kafka) • Support for Role Based Access Control (RBAC) • Remote Support – tools to collect logs and upload to IBM support • Backup / restore / DR of Metadata Database • Audit Trail – track and log user actions on Spectrum Discover • Dashboards to monitor Spectrum Discover health
Action Agent SDK Ecosystem	• SDK to help extend the platform capabilities to perform custom actions around data – data migration, archiving, content-based search & tagging, etc.
Classification, Tagging	• Apply custom tags based on system metadata, file system path or other criteria • Content-based classification & tagging based on the occurrence of user-definable keywords found in supported file types • Detect and classify several pre-defined types of PII and sensitive data • Use REGEX to define custom patterns for content-based search and tagging
GUI	• Basic Search • Advanced drill-down search • Dashboard to visualize storage consumption on a wide range of system and custom metadata • Create and schedule policies • Support for Role Based Access Control (RBAC)
Scalability	• Up to 100 billion indexed documents
Performance	• Ingest up over 1 Billion records per day
Quality	• Net Promoter Score (NPS) widget instrumented into Discover to gather NPS scores.

Licensing

Pricing Summary

- Licensed based on data managed by the Program (L-GMVS-BU26FM)
- Aggregate size of all the files that Spectrum Discover indexes and/or scans
- Ability to report on that size (whatever is indexed, logical size)
- For Spectrum Scale: can be configured to manage data on a specific file-system(s) or fileset(s) on that file-system.
- For Cloud Object Storage: can be configured to manage data on a specific vault(s).
- **90-day FREE trial**

Licensing Summary

- Licensed on a managed terabyte basis; flat pricing, no tiering; customers can manage as little or as much as they want.
- Orderable through either PPA or AAS; also available via eConfig for IBM Cloud Object Storage

IBM Spectrum Discover

New in 2.0.1

- Support for Heterogeneous Data Sources
- Content-based Keyword Search & Tagging
- Automatic classification of PII & sensitive data

Support for Heterogeneous Data Sources

FEATURE

NFS Support:

- Dell EMC Isilon
- NetApp

S3 Support:

- AWS S3
- Ceph

BENEFIT

With support for other popular 3rd party file and object storage systems, Spectrum Discover creates an open data ecosystem for unstructured data wherever it resides– on-premises or in the cloud.

Content-based Keyword Search & Tagging

FEATURE

Out-of-the-box support for content search enables end users to easily set up policies to automatically identify, classify and categorize data, which could be leveraged for specific business needs

BENEFITS

For the Data Scientist, CIO and the Data Analyst, the ability to curate, extract and gather data containing specific keywords is critical in large scale analytics involving vast amounts of unstructured data.

For the Data Steward and the CIO the ability to find and organize documents based on content greatly helps with their data administration efforts – for example, identifying data that may be subject to specific governance policies and/or compliance regulations.

Automatic classification of PII & sensitive data

FEATURE

Identifies key fields such as SSN, phone numbers, account numbers and many others to identify and tag content that contains PII & Sensitive Data.

BENEFIT

Automates the identification and classification of documents that could potentially contain Personally Identifiable Information (PII) and Sensitive Data.

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

Increase business value

Analytics

Uncover hidden data value

- Accelerate data identification for large-scale analytics
- Efficiently curate large-scale unstructured data and create custom datasets for AI / ML / Analytics workflows

Governance

Help mitigate risk / improve quality

- Automatically identify certain kinds of PII & sensitive data, and map this data to the right storage location
- Help reduce risk buried in unstructured data stores
- Tag / index data for eDiscovery & legal hold, helping speed up investigations

Optimization

Improve storage utilization

- Decrease storage CAPEX by facilitating data movement to colder, cheaper storage
- Increase storage efficiency by eliminating ROT data
- Reduce storage OPEX by improving storage administrator productivity

IBM Spectrum Discover

IBM Spectrum Discover Demo

- Ease of installation & configuration
- Highlights of new 2.0.1 functionality

IBM Spectrum Discover Redpaper

New IBM Redpaper now available

- **Provides a deep-dive, “how to” guide for IBM Spectrum Discover**
- **Includes new version 2.0.1 functionality**

Topics:

- **Overview of IBM Spectrum Discover**
- **Metadata Essentials**
- **Use Cases**
  - **Storage Optimization**
  - **Data Governance**
  - **Healthcare and Life Sciences**
- **Deep Inspection and AI Pipeline**
- **Installation and Setup**

Washington Systems Center - Storage

The Free IBM Storage Technical Webinar Series Continues in 2019...

Washington Systems Center – Storage experts cover a variety of technical topics.

Audience: Clients who have or are considering acquiring IBM Storage solutions. Business Partners and IBMers are also welcome.

To automatically receive announcements of upcoming Accelerate with IBM Storage webinars, Clients, Business Partners and IBMers are welcome to send an email request to accelerate-join@hursley.ibm.com.

Located in the Accelerate with IBM Storage Blog:
https://www.ibm.com/developerworks/mydeveloperworks/blogs/accelerate/?lang=en

Also, check out the WSC YouTube Channel here: https://www.youtube.com/channel/UCNuks0go01_ZrVVF1jgOD6Q

2019 Upcoming Webinars:

August 6 – IBM Storage SAN b-type Extension: Native IP vs FCIP
Register Here: https://ibm.webex.com/ibm/onstage/g.php?MTID=eba8b985837a2454480877deb0224114f

August 13 - Storage optimization with IBM Spectrum Discover
Register Here: https://ibm.webex.com/ibm/onstage/g.php?MTID=e9d8fdebce95d49d5148972c8b5cd967a

August 22 - LinuxONE Servers and IBM Storage Synergies
Register Here: https://ibm.webex.com/ibm/onstage/g.php?MTID=e246d7bbd6b0af257384b0e93c9032eec

August 29 - New content search capabilities in IBM Spectrum Discover
Register Here:https://ibm.webex.com/ibm/onstage/g.php?MTID=eea2e4a7977264229780d842beb9c7580