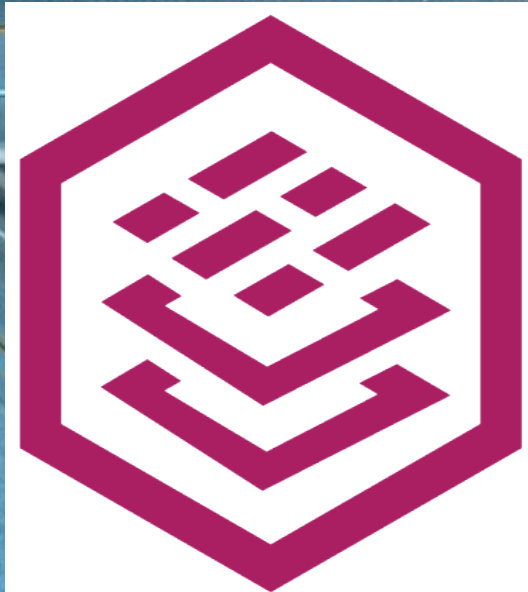




IBM Spectrum Discover

Exceptional Data Insight for Exabyte-Scale Unstructured Data Storage

Joseph Dain, Chief Architect, Spectrum Discover, STSM, Master Inventor

24 October 2018



- **Problem Statement**
- **Spectrum Discover Introduction and Value Proposition**
- **Architecture**
- **Use Cases and Demonstrations**
- **First Release Feature Highlights**
- **Free Trial and Useful Resources**



Biggest Unstructured Data Challenges



Source: Forrester Analytics, Global Business Technographics Data And Analytics Survey, 2017, Global Business Technographics Data And Analytics Survey, 2016 (Enterprises with 1000+ employees)



Unstructured Data is Hard to Manage

For Petabyte- and Exabyte-scale data stores...

- Challenging to pinpoint & activate relevant data for large-scale analytics
- Lack of fine-grained visibility needed to map data to business priorities
- Difficult to remove redundant, trivial & obsolete data
- Tough to identify & classify sensitive data



Metadata is the Key to Data Insight

- Metadata is the structured data about the unstructured object
 - Who, what, when, where, and why of account, container, object, stream, dir, file
 - Perfect for indexing and searching
- Metadata may be separate from the data, stored with the data, or derived from the data
 - Posix inode plus extended attributes
 - Standard document headers (doc, ppt, mp3, dicom, pdf, jpeg, GeoTIFF)
 - Custom metadata tags
 - AI derived metadata

System Metadata

Location

Size

Owner

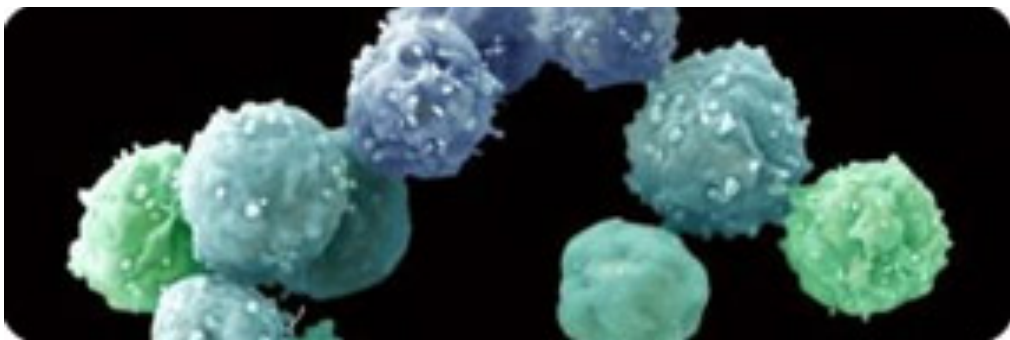
Group

Permissions

Last-Modified

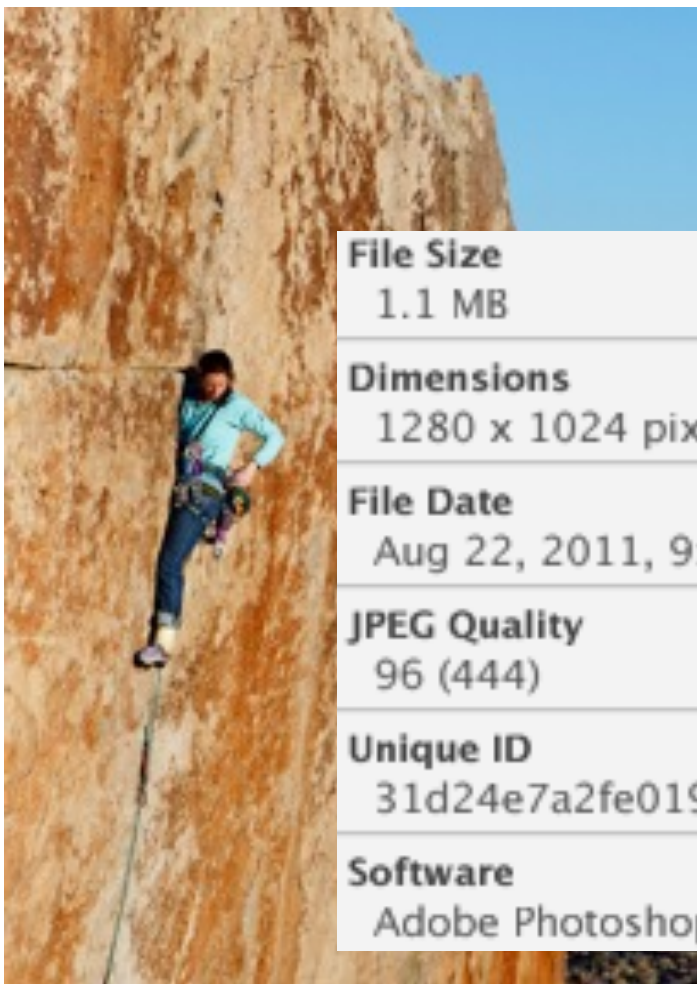
...

Biomedical



Age, Biomarkers, Developmental Stage,
Cell Surface, Markers, Cell Type/Cell Line,
Disease State, Extract Molecule, Genetic
Characteristics, Immunoprecipitation,
antibody, Organism,

Image



File Size

1.1 MB

Dimensions

1280 x 1024 pixels

File Date

Aug 22, 2011, 9:42 AM

JPEG Quality

96 (444)

Unique ID

31d24e7a2fe01906000000000000000

Software

Adobe Photoshop CS5 Macintosh



Natural Language Processing

2004-49-188A.txt

1

X1 a 1999 **blue** **Carry**, was traveling southbound in the second **lane** of a four-lane divided (seven **lanes** overall, divided by raised median), concrete **roadway** approaching an **intersection**

2

X2 a 2004 Mercedes S430, was northbound in the fourth **lane** of a four-lane, divided (seven **lanes** overall, divided by raised median), concrete **roadway** about to turn left into westbound traffic at the same **intersection**

3

As both **vehicles** entered the **intersection** the **front** of X1 impacted the **rear** of X2

4

X1 rotated clockwise as X2 rotated counter-clockwise, and the left side of X1 impacted the right side of X2 in a sideslap configuration.

5

Both **vehicles** moved southwest to final rest.

6

Both **vehicles** were towed due to damage.

7

The unrestrained **driver** of X1 was hospitalized with foot and rib fractures as well as a liver laceration.

Entity

Identify

Type

Subtype

Id

ACCIDENT_CAUSE

ACCIDENT_OUTCOME

CONDITION

IMPACT

MANUFACTURER

MODEL

MODEL_YEAR

PART_OF_CAR

PERSON

STRUCTURE

VEHICLE

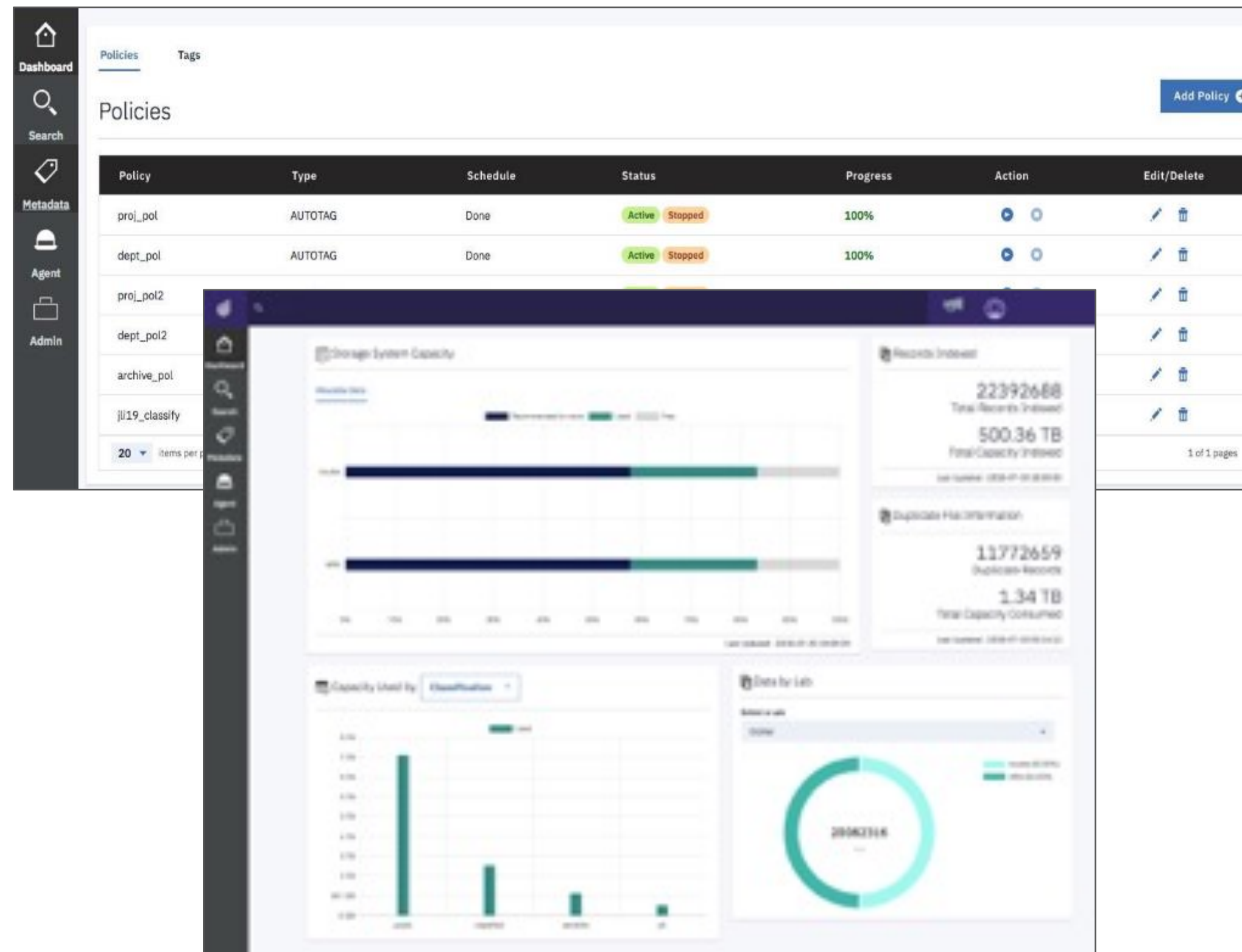
PYTORCH



Introducing



IBM Spectrum Discover



Data Insight for Analytics, Governance and Optimization

- **Automate** the 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 and data consumer productivity
- **Leverage extensibility** using the API, custom tags and policy-based workflows to orchestrate content inspection and activate data in AI, ML and analytics workflows

Why IBM Spectrum Discover?

It provides a platform enabling organizations to . . .

Analyze

Uncover Hidden Data Value

- Accelerate data identification for large-scale analytics
- Operationalize tasks to reduce the burden of data preparation
- Orchestrate ML/DL and MapReduce processes



Govern

Mitigate Risk and Improve Data Quality

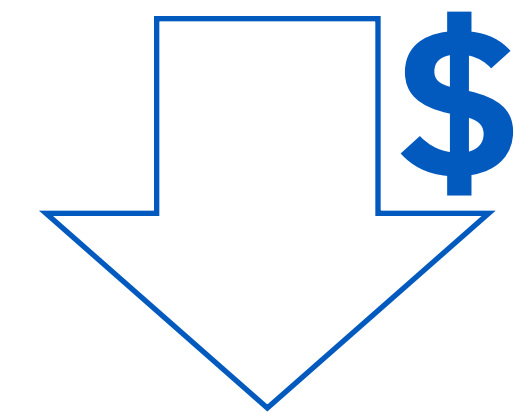
- Ensure data adheres to governance policies
- Reduce the risk buried in unstructured data stores
- Speed investigation into legal discover and/or regulatory audit



Optimize

Improve Storage Utilization

- Decrease storage capex by facilitating data movement to colder, cheaper storage
- Increase storage efficiency by eliminating redundant and ROT data
- Reduce storage opex by improving storage administrator productivity



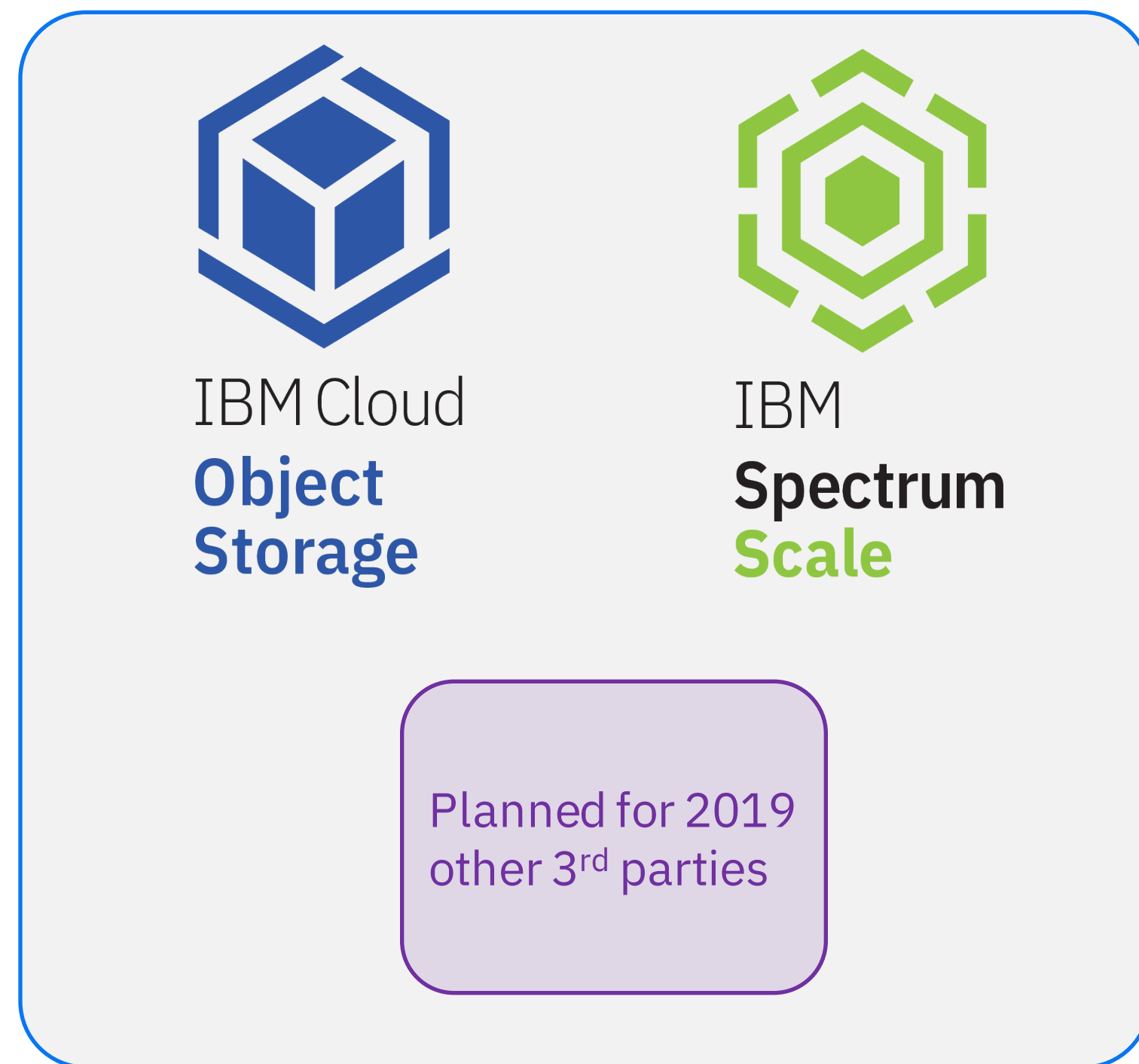
Product Architecture



IBM Spectrum Discover Architecture

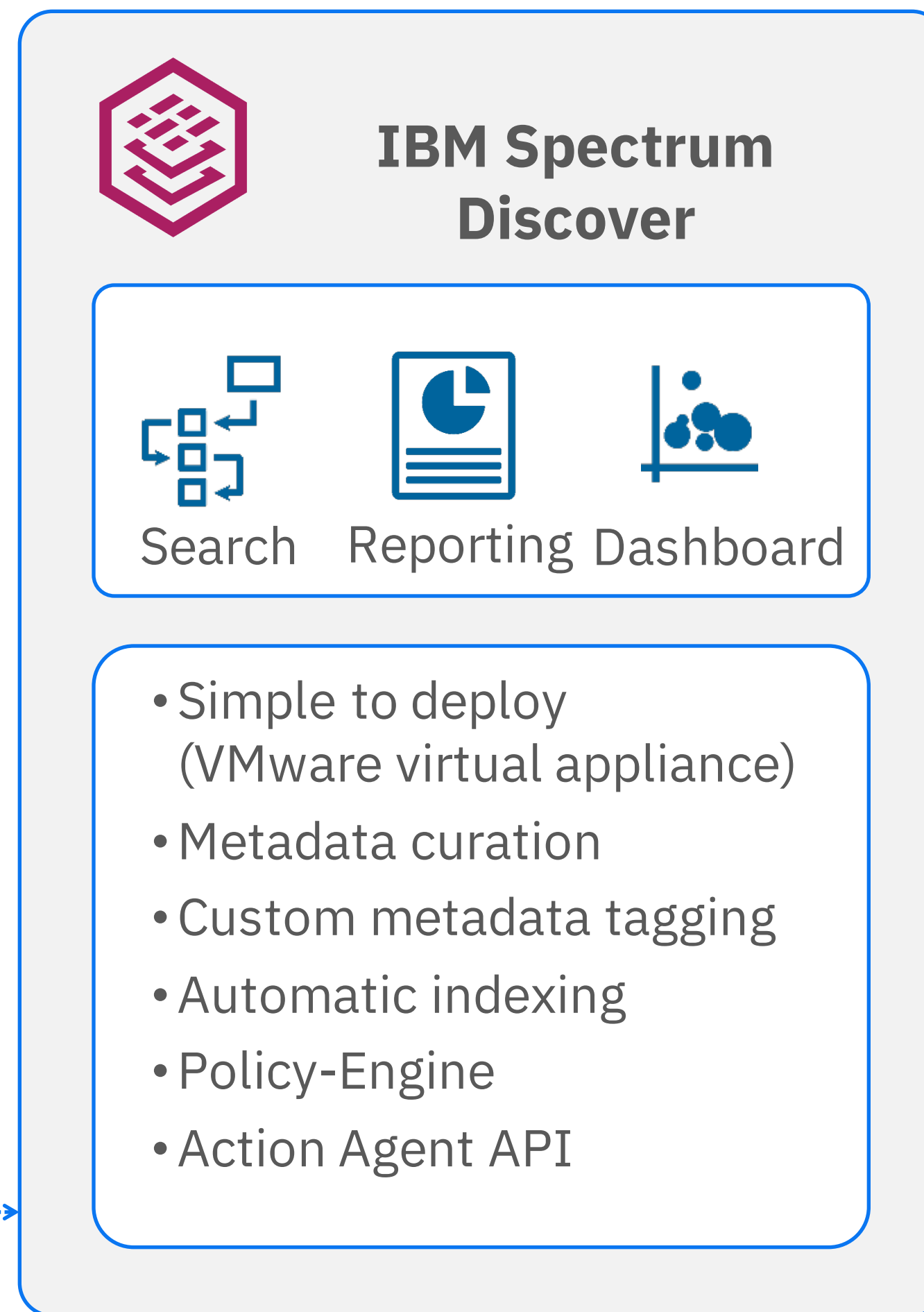


File and Object Storage

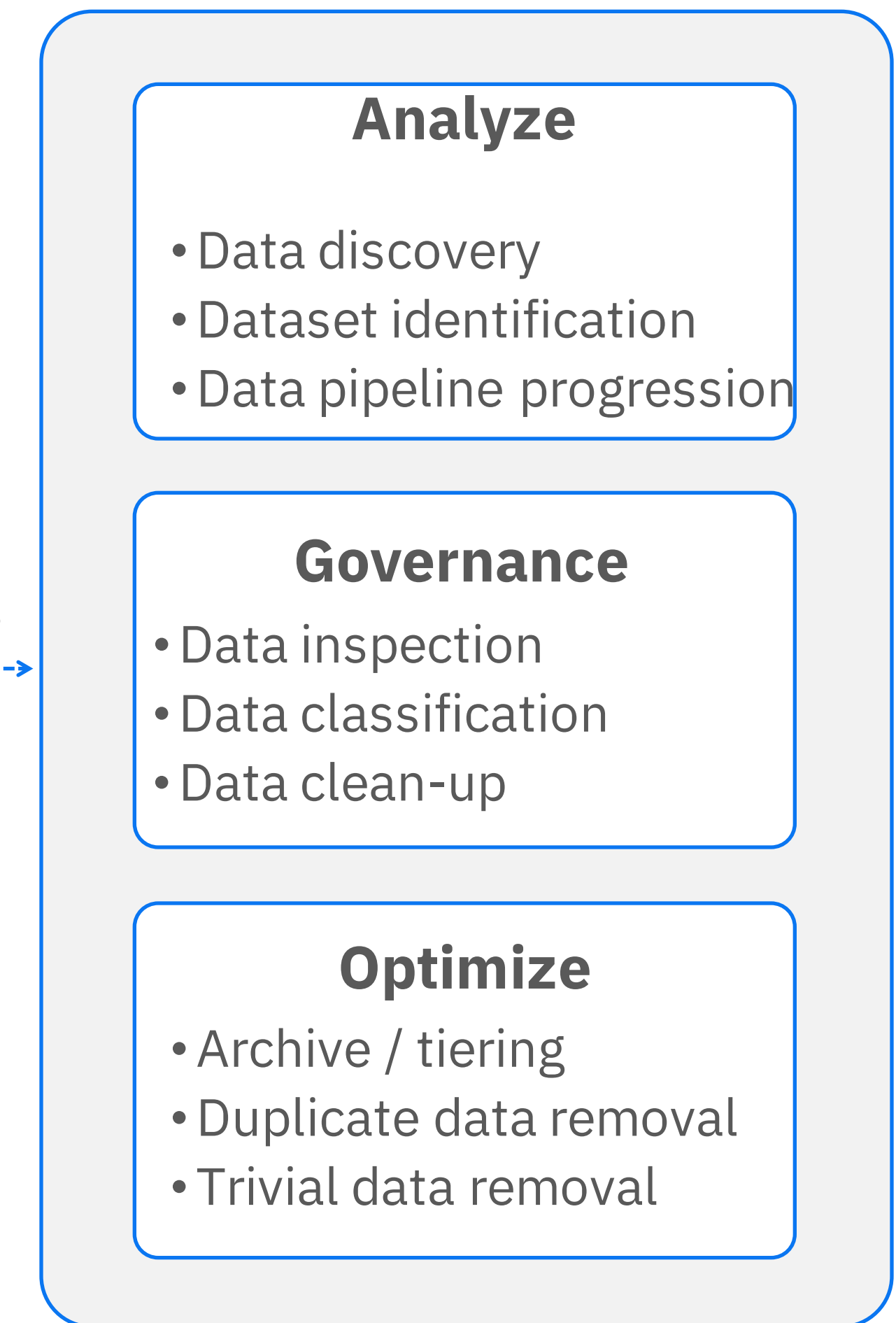


Scanning and
Event Notifications

Data Insight



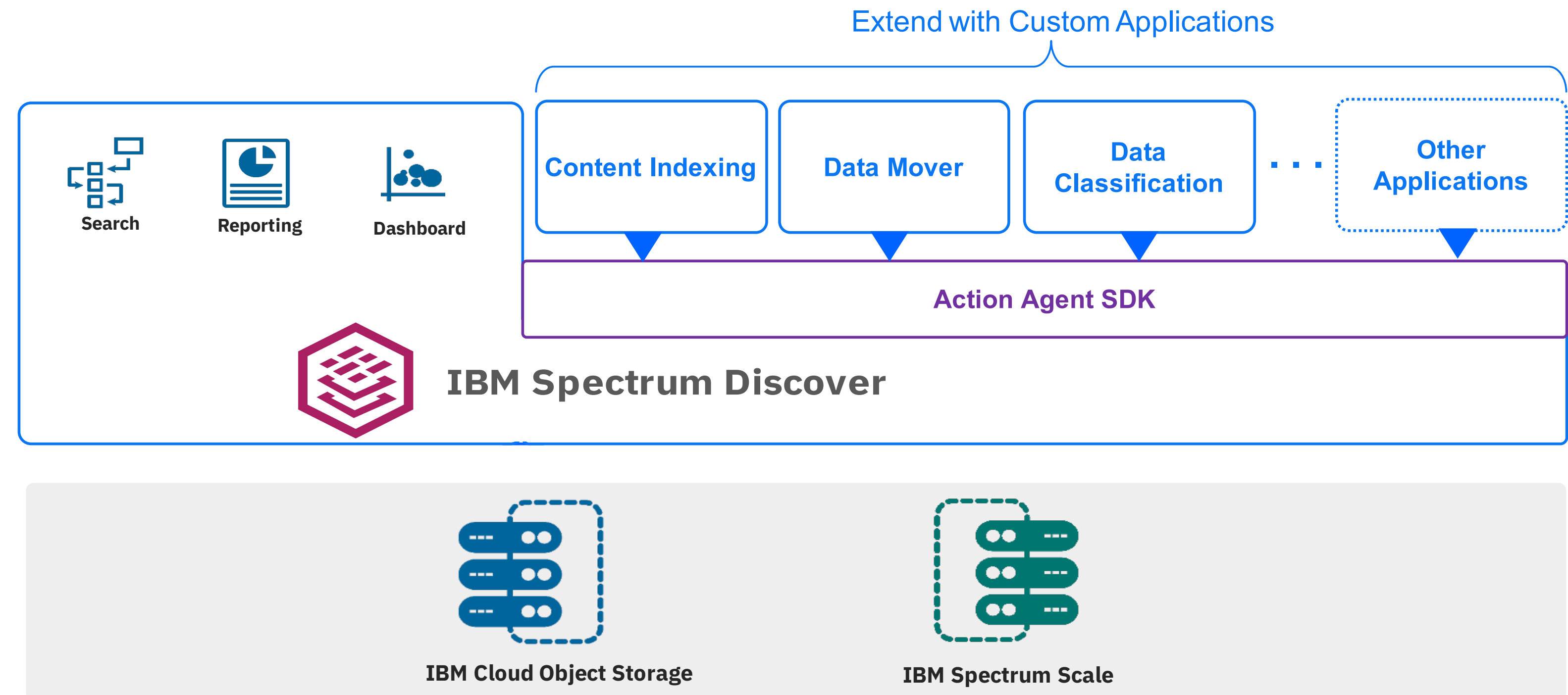
Data Activation/Optimization



Extensible Foundation for Data Insight



- Action Agent SDK extends capabilities via well defined API
- Customize actions taken based on Discover metadata
 - ❖ Content indexing
 - ❖ Data movement (tiering)
 - ❖ Classification
 - ❖ Sensitive data identification
 - ❖ ROT Detection/Disposal
 - ❖ Etc...
- Integrate with upstream information management applications



An aerial, high-angle photograph of a large, modern skate park. The park is constructed from light-colored concrete with various curved ramps, bowls, and ledges. Numerous people are scattered throughout the park, engaged in activities like skateboarding, longboarding, and riding bicycles. Some individuals are in motion, creating slight blurs, while others are stationary. The overall scene is active and dynamic.

Use Cases

IBM

Optimize: Place Data on Appropriate Storage Tier



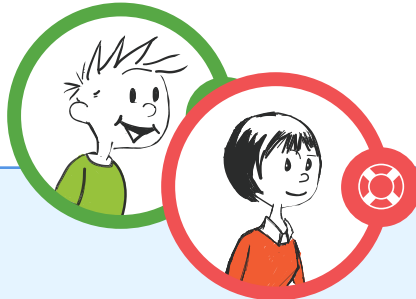
Business Leaders

(Dir / VP of IT / CIO / CFO)



IT Infrastructure Team

(Storage Administrator)



Data Owners

(Researcher / Data Scientist / Lab Manager / RIM)



Risk

- Ensuring new Business Strategies are not delayed by lack of capacity.
- Ensuring existing applications have capacity for expansion.

Costs

- Unnecessary purchases due to inefficient management of storage.
- Avoid unplanned Operational Expense due to business processes not being followed.

Before

Managing Capacity Utilization

- Monitor multiple vendor technologies.
- Wait for the signs of quota threshold violations.
- Track down Data Owners and ask them to audit their data placement, retention, and take corrective action.
- Wait for reply from Data Owner regarding what data can be archived.

Before

Manually Managing the Data

- Track data ownership in a spreadsheet
- Relies on peers following best-practices in order to track data.
- Tries to figure out what data can be archived in order to save on storage costs.
- Open ticket to have data archived and migrated.

with Spectrum Discover

Shared Management Responsibility

- Visualize/monitor utilization of multiple storage sources with drill-down analytics.
- Custom, policy-driven data tagging identifies candidate files for action.
- Intelligently manage data placement on appropriate storage tier (archive)

with Spectrum Discover

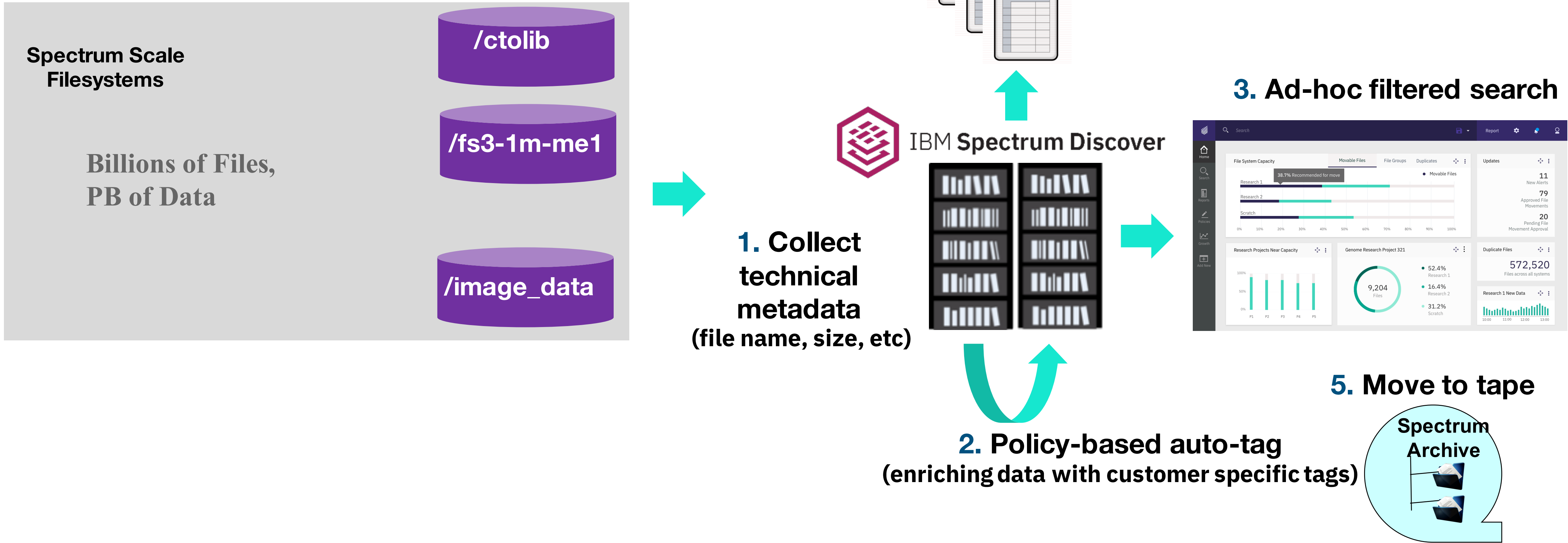
Shared Management Responsibility

- Custom, policy-driven data tagging identifies candidate files for action.
- Multi-faceted search to identify other candidates for archiving.
- Improved understanding of data use profile



IBM Spectrum Discover

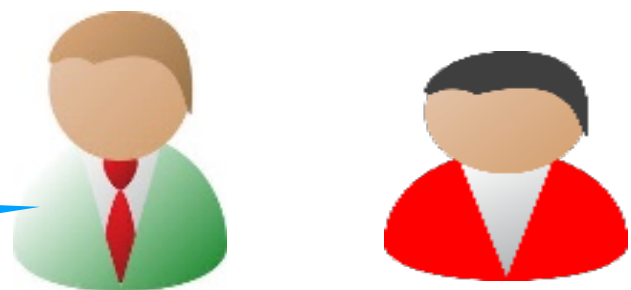
Use Case: Curating the Research Data for Placement Optimization



Use Case: PHI and non-PHI Classification of DICOM Images



Show me where my PHI data resides



IBM ESS




IBM Spectrum Scale

System Metadata Events

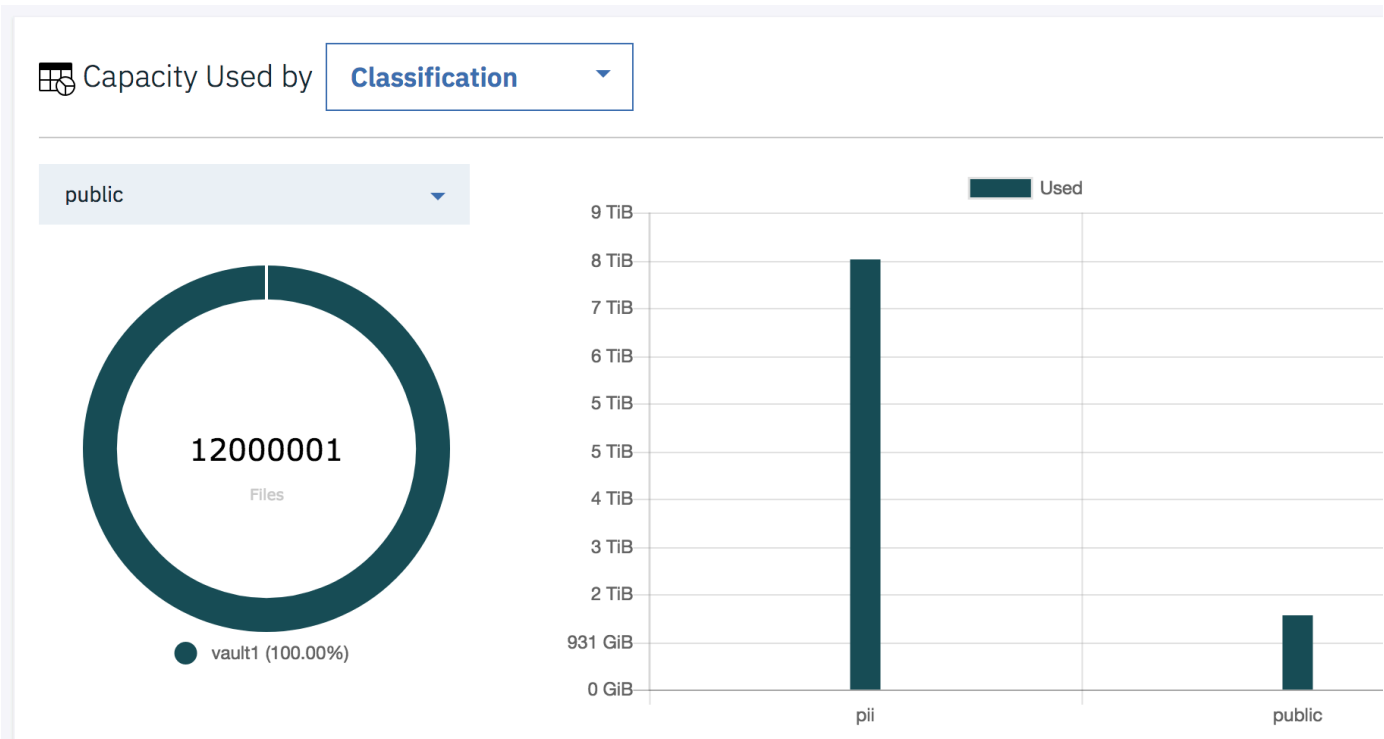
IBM Spectrum Discover

Classification:

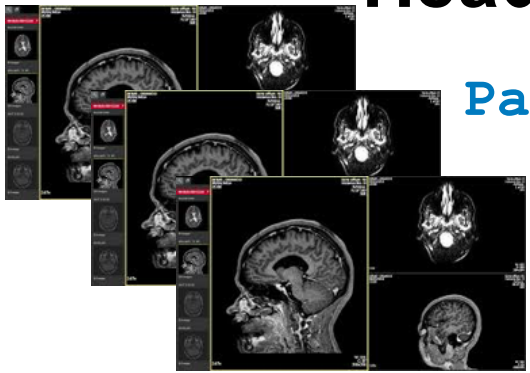
PHI

NON-PHI

Policy Engine:
- Header Extraction



DICOM Header Extraction (pydicom)



PatientName: Pikachu



Analyze: Automate data labeling for large-scale Analytics



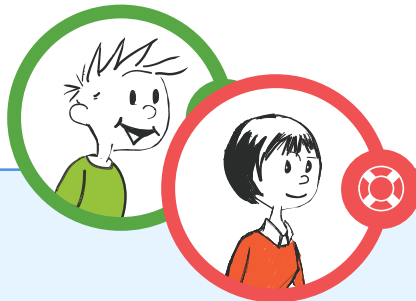
Business Leaders

(Dir / VP of IT / CIO / CFO)



IT Infrastructure Team

(Storage Administrator)



Data Owners

(Researcher / Data Scientist / Lab Manager / RIM)



Risk

- Ensuring the use of highly relevant data while conducting AI/ML experiments.
- Inability to reproduce key scientific results because relevant data could not be identified.

Costs

- Wasting valuable time and resources in the data preparation stage.
- Increased time-to-market for key products and losing competitive advantage.

Before

Want to help identify your data, but ...

- Filtering records based on data owners, dates or other system metadata is insufficient.
- When data is coming in at high volumes and high speeds from multiple sources, no easy way to label and track this data.

Before

Dig through the weeds

- Request peers to follow best practices to manage data used for key AI/ML experiments.
- Ask peers to maintain Excel sheets to keep track of data used for different projects
- Unable to verify accuracy and reliability of manually tracked data.

with Spectrum Discover

Custom and automated labeling ...

- Custom, policy-driven data tagging maps data to projects, departments, business units and other relevant entities.
- Automate the labeling of high-volume data coming from multiple data sources.

with Spectrum Discover

Less time searching for data, more time developing models ...

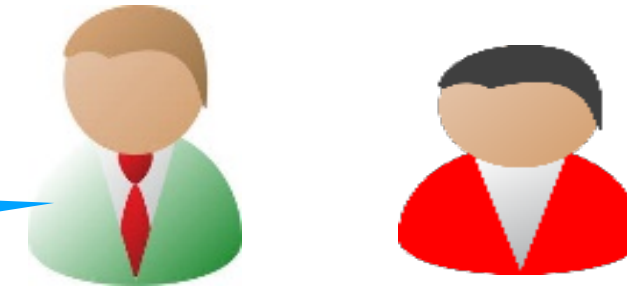
- Easily identify data used for past experiments to continue or reproduce those experiments.
- Leverage high-volume data from multiple sources in real time, reduce time-to-market and gain competitive advantage.



IBM Spectrum Discover

Use Case: Curating the Data for Researchers and Data Scientists

Data Scientists /
researchers



Give me medical images
of males age 45-60



Using Spectrum Discover, Data
Scientists can:

- ✓ Add custom tags to files to match project, department, data owner, etc.
- ✓ Use deep inspection API's to index specific content in the Spectrum Discover database.

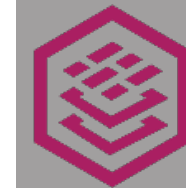
Enabling other researchers to later:

- ✓ Efficiently search and extract information highly relevant to their projects from Petabytes of data.
- ✓ Find similar projects/experiments in the past.
- ✓ Find other researchers with similar interests for potential collaboration.

IBM ESS



System
Metadata Events



IBM Spectrum Discover

Classification:

Gender

DOB

Policy Engine:



- Header
Extraction

DICOM



Header Extraction

Gender: Male
DOB: 19490301

Governance: Eliminate Redundant, Obsolete, Trivial Data



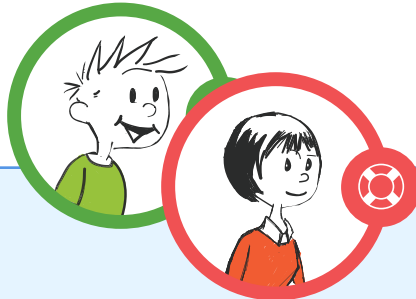
Business Leaders

(Dir / VP of IT / CIO / CFO)



IT Infrastructure Team

(Storage Administrator)



Data Owners

(Researcher / Data Scientist / Lab Manager / RIM)



Risk

- Ensuring new Business Strategies are not delayed by lack of capacity.
- Ensuring existing applications have capacity for expansion.

Costs

- Unnecessary purchases due to inefficient management of storage.
- Avoid unplanned Operational Expense due to business processes not being followed.

Before

Hope, but can't confirm

- Can't get a good look into the files stored in the storage environment.
- Hopefully, Data Owners are trying to reduce duplication since they are paying for every copy. (And so are we.)
- Trust that Data Owners delete obsolete and trivial data from corporate storage resources

Before

Dig through the weeds

- Request peers to follow best practices to manage data and storage resources.
- Ask peers to look through their data and delete unnecessary data and duplicates...
- ...Or try and look through all the data myself.

with Spectrum Discover

Custom and Automated Detection

- Custom, policy-driven data tagging identifies obsolete and trivial data
- Automate detection of potentially duplicate files with the same name and same size across the environment.
- View details of the duplicate set or file-level details.

with Spectrum Discover

Focused Discussions

- Collaborate with other data owners to decide which copies of the data are necessary.
- Delete duplicate data or flag unneeded copies for archive.



IBM Spectrum Discover

Key Features



Connectors	• Scanner for IBM Cloud Object Storage (COS) and IBM Spectrum Scale • 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 • 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.
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)
Performance	• Ingest over 1 Billion records per day

An aerial, high-angle photograph of a large, modern skate park. The park is constructed from light-colored concrete with various ramps, bowls, and ledges. Numerous people are scattered throughout the park, some standing, some riding skateboards, and others on bicycles. The perspective is from directly above, looking down on the scene. The overall tone is bright and active.

Resources, Free Trial, and Spectrum Scale Live Events Tech Preview



Free Trial Software Download



- Available at GA (Planned 11/2/2018)
- 90 Day Free Trial
 - At end of 90 days, code no longer accessible by client w/o approved extension or purchase of full license
- Full function version of code
 - Not limited scale or function set
 - At termination of trial, access terminates
- Restriction(s)
 - Cannot upgrade from single node trial to multi-node production
- Support for trial: spdiscov@us.ibm.com (available at GA)

Spectrum Discover Product Page



IBM
Spectrum
Discover

IBM

Marketplace ▾

Search IBM Marketplace

🔍

IBM Spectrum Discover

Overview

Details

Contact us

IT infrastructure > Storage > Spectrum storage >

IBM Spectrum Discover

Provides modern metadata management software for petabyte-scale file and object storage

Contact us for pricing

Learn how IBM is innovating storage and data management from creation to archive. → Watch the webcast



THANK YOU!

IBM Global Financing offerings are provided through IBM subsidiaries and divisions worldwide to qualified commercial and government clients. IBM Global Financing lease and financing offerings are provided in the United States through IBM Credit LLC. Rates and availability are based on a client's credit rating, financing terms, offering type, equipment and product type and options, and may vary by country. Non-hardware items must be one-time, non-recurring charges and are financed by means of loans. Other restrictions may apply. Rates and offerings are subject to change, extension or withdrawal without notice and may not be available in all countries. IBM and IBM Global Financing do not, nor intend to, offer or provide accounting, tax or legal advice to clients. Clients should consult with their own financial, tax and legal advisors. Any tax or accounting treatment decisions made by or on behalf of the client are the sole responsibility of the client. For IBM Credit LLC in California: Loans made or arranged pursuant to a California Financing Law license.

For more information, visit: ibm.com/financing

© Copyright IBM Corporation 2018.

IBM Cloud – Middleware, New Orchard Road Armonk, NY 10504.

Produced in the United States of America, August 2017.

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 "Copyright and trademark information" at <http://www.ibm.com/legal/copytrade.shtml>. This document is current as of the initial date of publication and may be changed by IBM at any time. Not all offerings are available in every country in which IBM operates. The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions. THE INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS" WITHOUT ANY WARRANTY, EXPRESS OR IMPLIED, INCLUDING WITHOUT ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND ANY WARRANTY OR CONDITION OF NON-INFRINGEMENT. IBM products are warranted according to the terms and conditions of the agreements under which they are provided. The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

