Introduction to Spectrum Discover

Spectrum Scale German User Meeting 2022
Cologne, Germany – October 19-20, 2022

Lars Lauber (IBM)
lauberla@de.ibm.com

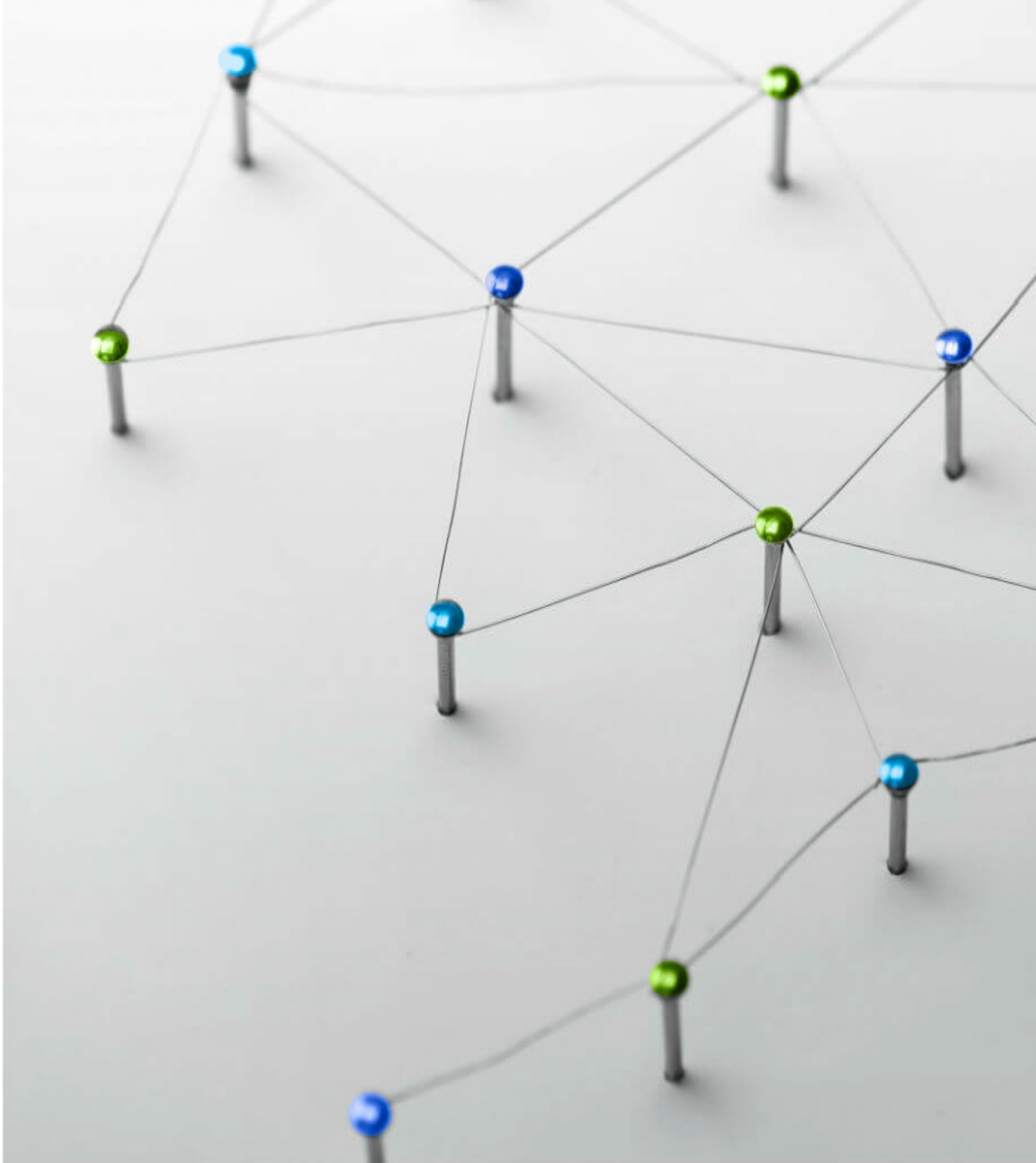
Disclaimer

IBM's statements regarding its plans, directions, and intent are subject to change or withdrawal without notice at IBM's sole discretion. Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision. The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

IBM reserves the right to change product specifications and offerings at any time without notice. This publication could include technical inaccuracies or typographical errors. References herein to IBM products and services do not imply that IBM intends to make them available in all countries.

Content

- The struggle with AI and big data
- Metadata holds the key for unstructured data
- What can Spectrum Discover do to help?
- Using Spectrum Discover for data analysis
- Spectrum Discover Integration with Spectrum Scale



Those who can harness the power of their data have a competitive advantage!

20%

of the world's data is searchable and anybody can get to it

The other

80%

is like gold

"Initial data analysis for the simplest project requires brutal, excruciating data wrangling. More than 80% of a project's time spent is me trying to **parse these files** and **cross data sources** in order to build viable datasets."

– Data Scientist at a large company

https://datascience.stackexchange.com/questions/48531/how-much-of-data-wrangling-is-a-data-scientists-job

Existing data is difficult to leverage and trapped

50%

want to use long term data for AI/Analytics

84%

of digital transformation projects fail due to siloed data and unreliable integration

- Complex to manage
- Difficult to leverage data from silos
- Not designed to leverage AI
- Inefficient to search

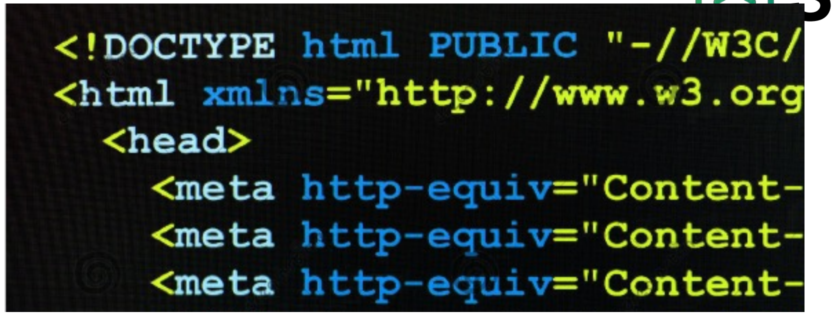
https://www.bitpipe.com/detail/RES/1546964826_733.html

AI & big data *can* be challenging …

- ❑ Data volume and data quality
- ❑ Difficult to combine multiple data sources
- ❑ Skills gaps to analyze data
- ❑ Current infrastructure complexity
- ❑ Where and how to begin?

… but they don’t have to be!

Metadata is the Key to Data Organization & Insight



```
<!DOCTYPE html PUBLIC "-//W3C/
<html xmlns="http://www.w3.org
  <head>
    <meta http-equiv="Content-
    <meta http-equiv="Content-
    <meta http-equiv="Content-
```

Three Types of Metadata

1

Metadata can come from a system

(owner, last modified, size, type, etc.)

2

Metadata can be custom

(map to various business & scientific aspects)

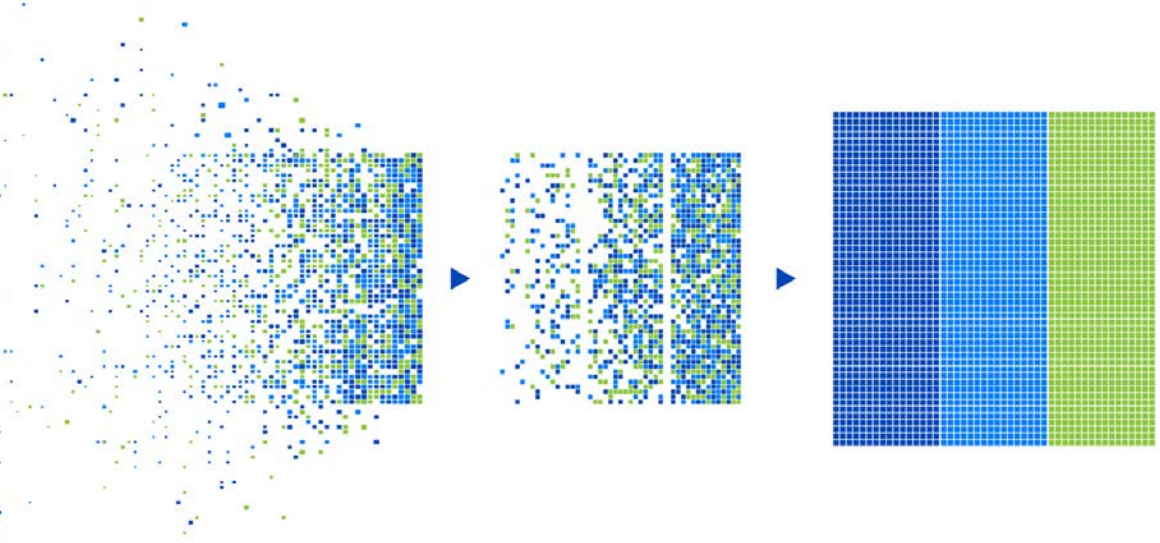
3

Metadata can be derived from analytics

(percent confident)

Metadata is the key

Bring structure to unstructured data …



Why metadata?

- Improves management
- Organizes large amounts of data
- Creates “indexes” or “tags” for data
- Provides a fast way to search and analyze

Metadata is data about data

- Context for data classification and 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 and manage** assets that add value
- **Simplify search & access** to critical data
- **Define and execute policies** based on metadata
- **Improve** time-to-value and storage economics
- **Enrich and increase** the value of data

IBM Spectrum Discover

- **Data insights:**
  Multi-vendor connections

- **Better AI:**
  Search billions of records < second

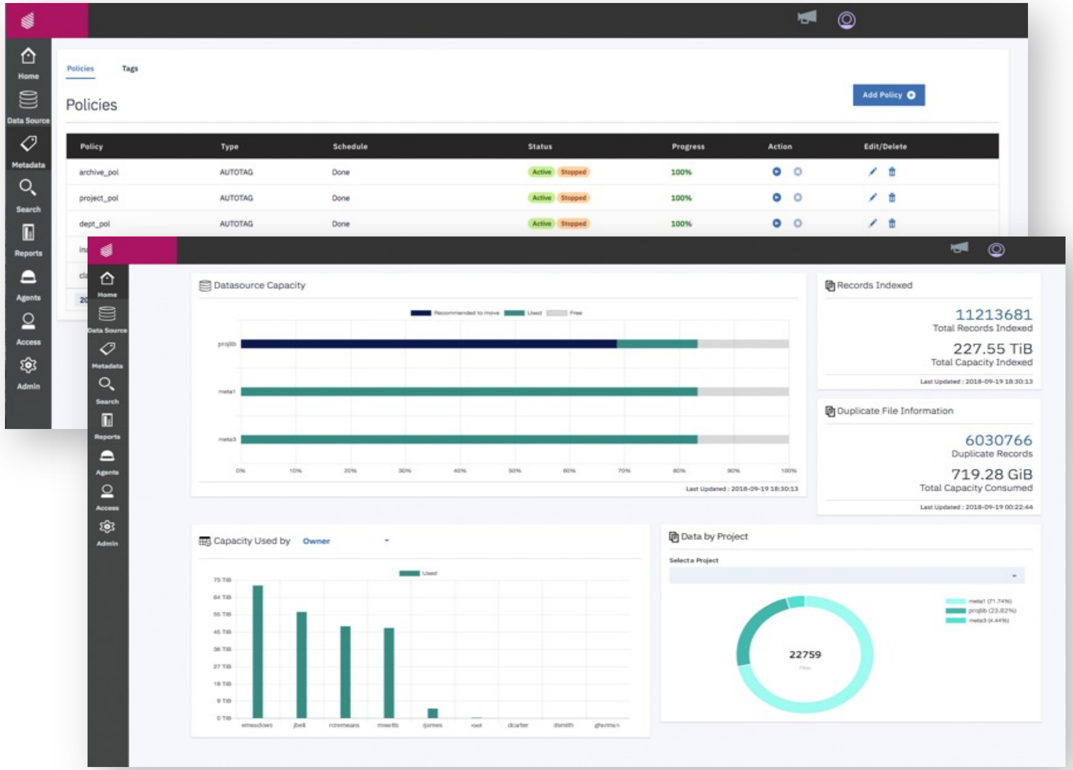
- **Optimize data workflows:**
  Policy based automation and auto tag data

- **Data security and compliance:**
  Discover security, compliance and governance issues before they become problematic

- **Business Value:**
  Link data to IBM Watson solutions and
  IBM Cloud Pak for Data

Comprehensive search and AI analysis with a data catalog and policy engine

IBM Spectrum Discover



Policies

Policy	Type	Schedule	Status	Progress	Action	Edit/Delete
archive_pol	AUTOTAG	Done	Active Stopped	100%		
project_pol	AUTOTAG	Done	Active Stopped	100%		
dept_pol	AUTOTAG	Done	Active Stopped	100%		

Datasource Capacity

Last Updated : 2018-09-19 18:30:13

Records Indexed

11213681
Total Records Indexed

227.55 TiB
Total Capacity Indexed

Last Updated : 2018-09-19 18:30:13

Duplicate File Information

6030766
Duplicate Records

719.28 GiB
Total Capacity Consumed

Last Updated : 2018-09-19 00:22:44

Capacity Used by Owner

Data by Project

Select a Project

IBM Spectrum Discover overview

Where

Backup, File & Object Storage

IBM **Spectrum Archive**

IBM **Spectrum Scale**

IBM Cloud **Object Storage**

IBM **Spectrum Protect**

IBM Elastic Storage Server

ISILON

amazon web services™ S3

NetApp

ceph

What

Index & Tag Big Data

IBM **Spectrum Discover**

- Simple to deploy (VMware virtual appliance)
- Metadata curation
- Custom metadata tagging
- Automatic indexing
- Content inspection
- Policy-Engine
- Application plugin API / SDK

Why

High Speed Data Insight

Large-Scale Analytics and AI/ML
- Data discovery
- Dataset identification
- Data pipeline progression

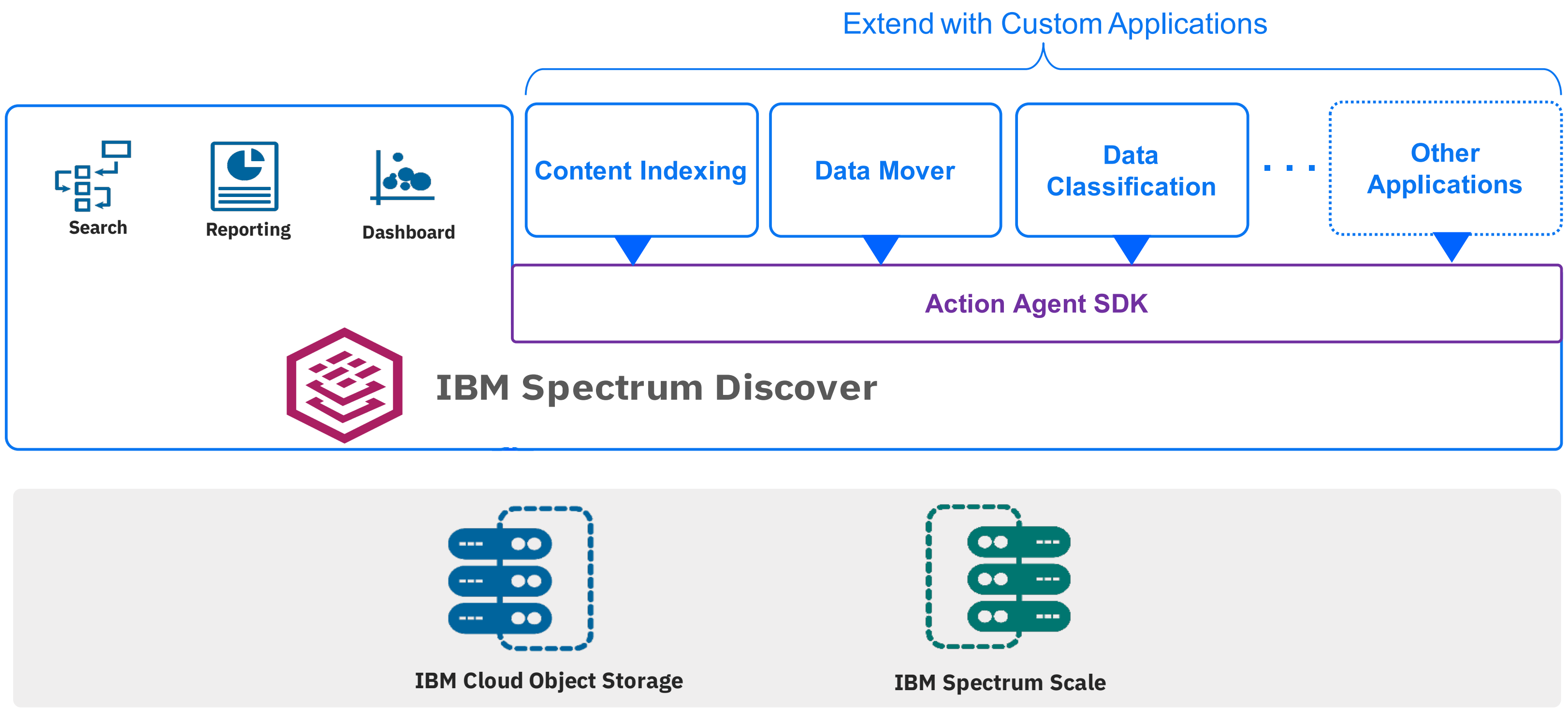
Data Governance
- Data inspection and classification
- Data clean-up

Data Optimization
- Archive / tiering
- Duplicate data removal
- Trivial data removal

Data Management
- Automate Tags for custom insight
- Create reports or directly search data
- Search content for fast discovery

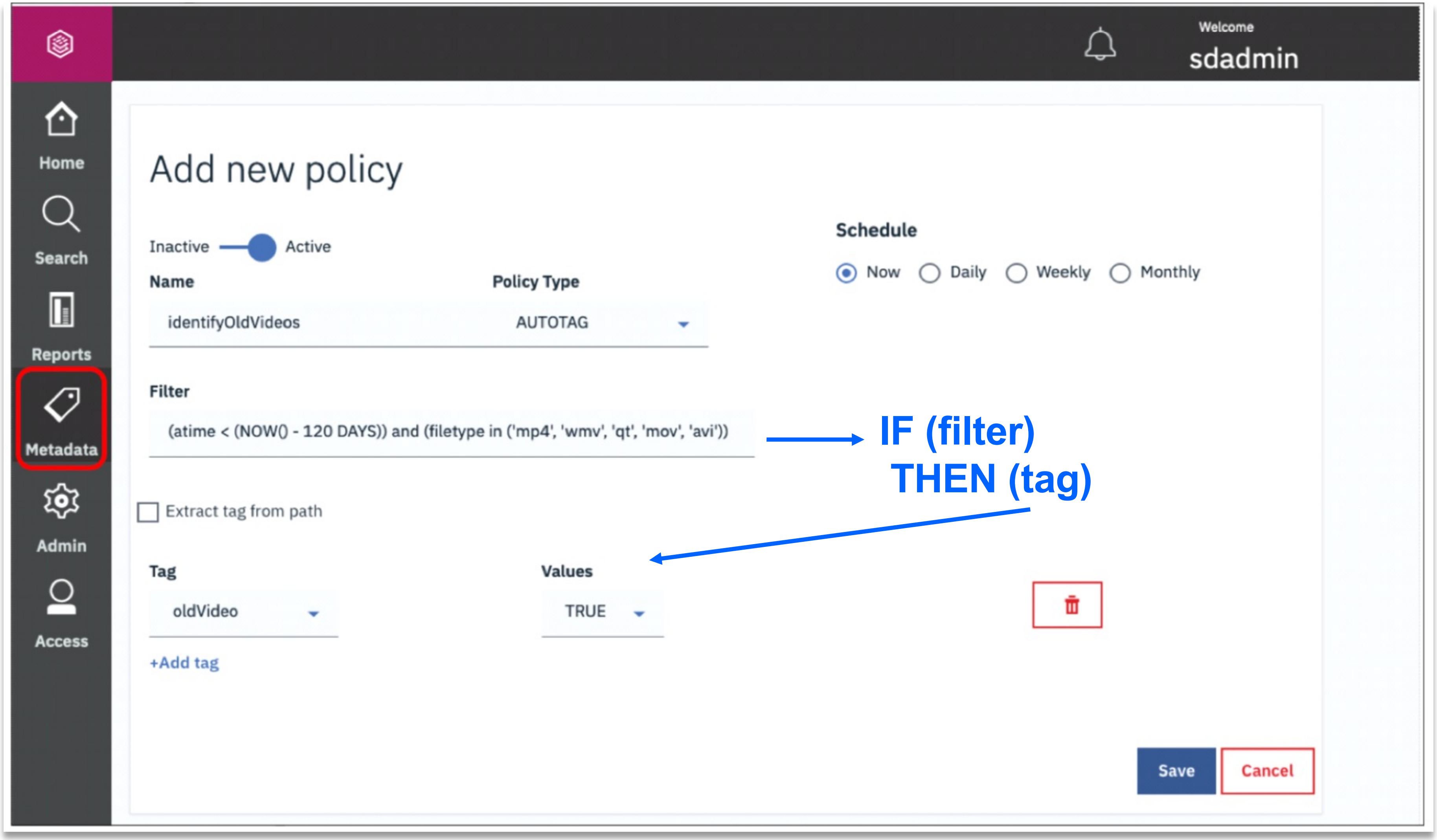
Extensible Foundation for Data Insight

- **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**



Easily create a policy to “tag” items based on a filter

Data Mapping



Add new policy

Inactive ● Active

Name	Policy Type
identifyOldVideos	AUTOTAG

Schedule

● Now ○ Daily ○ Weekly ○ Monthly

Filter

(atime < (NOW() - 120 DAYS)) and (filetype in ('mp4', 'wmv', 'qt', 'mov', 'avi'))

□ Extract tag from path

Tag	Values
oldVideo	TRUE

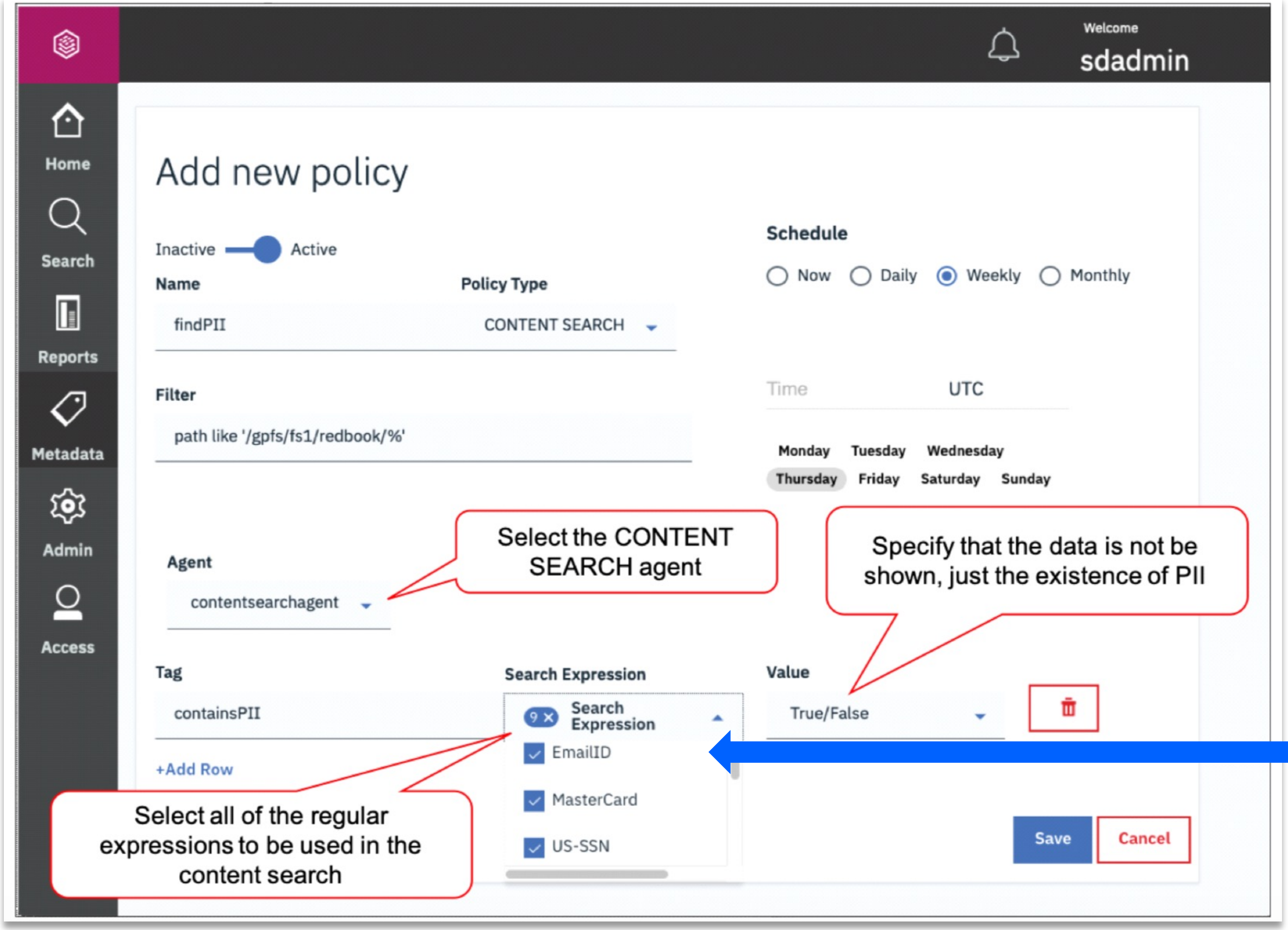
+Add tag

Save Cancel

IF (filter)
THEN (tag)

Search inside files/objects to find patterns and create new metadata tags

Data Mapping



Add new policy

Inactive ● Active

Name: findPII

Policy Type: CONTENT SEARCH

Filter: path like '/gpfs/fs1/redbook/%'

Schedule: ○ Now ○ Daily ● Weekly ○ Monthly

Time UTC

Monday Tuesday Wednesday Thursday Friday Saturday Sunday

Agent: contentsearchagent

Tag: containsPII

Search Expression: 9 × Search Expression

☑ EmailID

☑ MasterCard

☑ US-SSN

Value: True/False

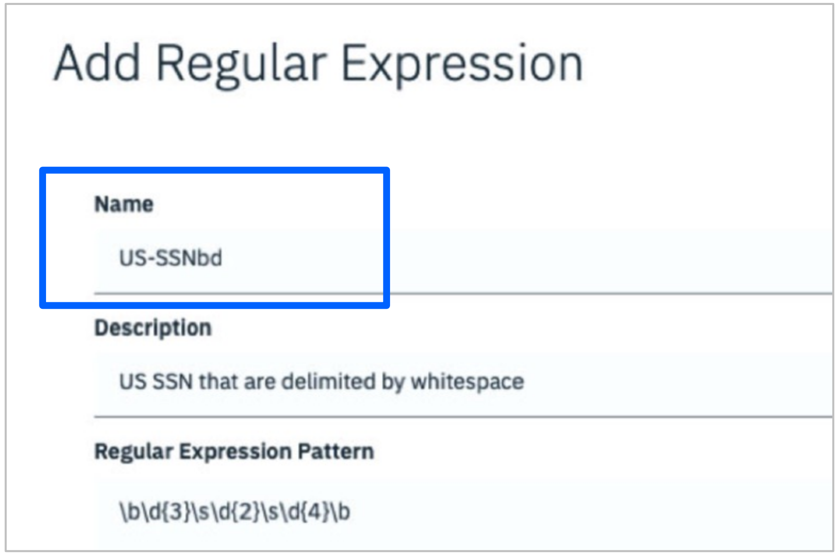
+Add Row

Save Cancel

Select the CONTENT SEARCH agent

Specify that the data is not be shown, just the existence of PII

Select all of the regular expressions to be used in the content search



Add Regular Expression

Name: US-SSNbd

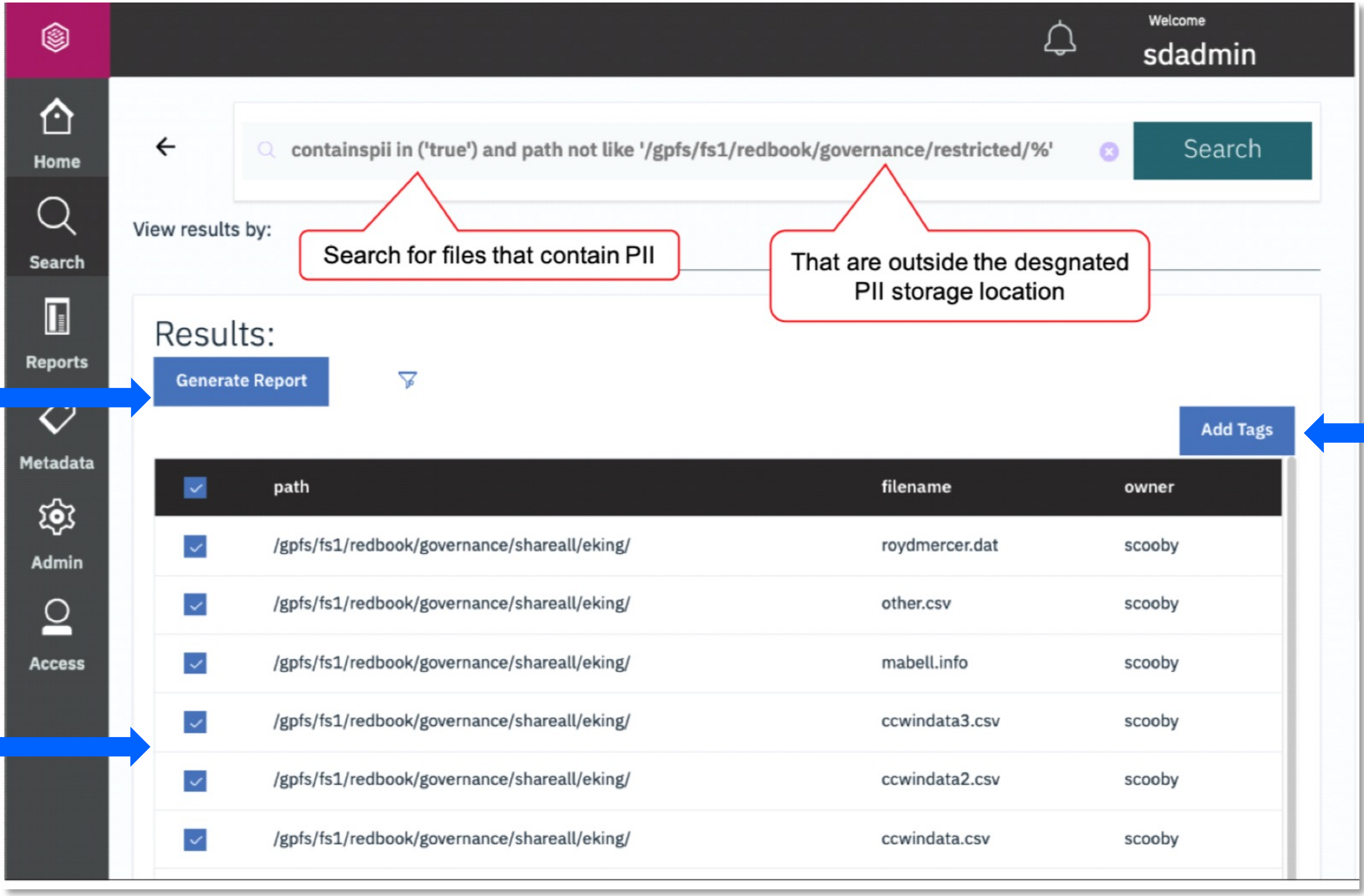
Description: US SSN that are delimited by whitespace

Regular Expression Pattern: `\b\d{3}\s\d{2}\s\d{4}\b`

Easily create your own custom search patterns

Discover your data with simple interface or report generation

Data Visualization



containspii in ('true') and path not like '/gpfs/fs1/redbook/governance/restricted/%'

Search for files that contain PII

That are outside the desgnated PII storage location

View results by:

Results:

☑	path	filename	owner
☑	/gpfs/fs1/redbook/governance/shareall/eking/	roydmercer.dat	scooby
☑	/gpfs/fs1/redbook/governance/shareall/eking/	other.csv	scooby
☑	/gpfs/fs1/redbook/governance/shareall/eking/	mabell.info	scooby
☑	/gpfs/fs1/redbook/governance/shareall/eking/	ccwindata3.csv	scooby
☑	/gpfs/fs1/redbook/governance/shareall/eking/	ccwindata2.csv	scooby
☑	/gpfs/fs1/redbook/governance/shareall/eking/	ccwindata.csv	scooby

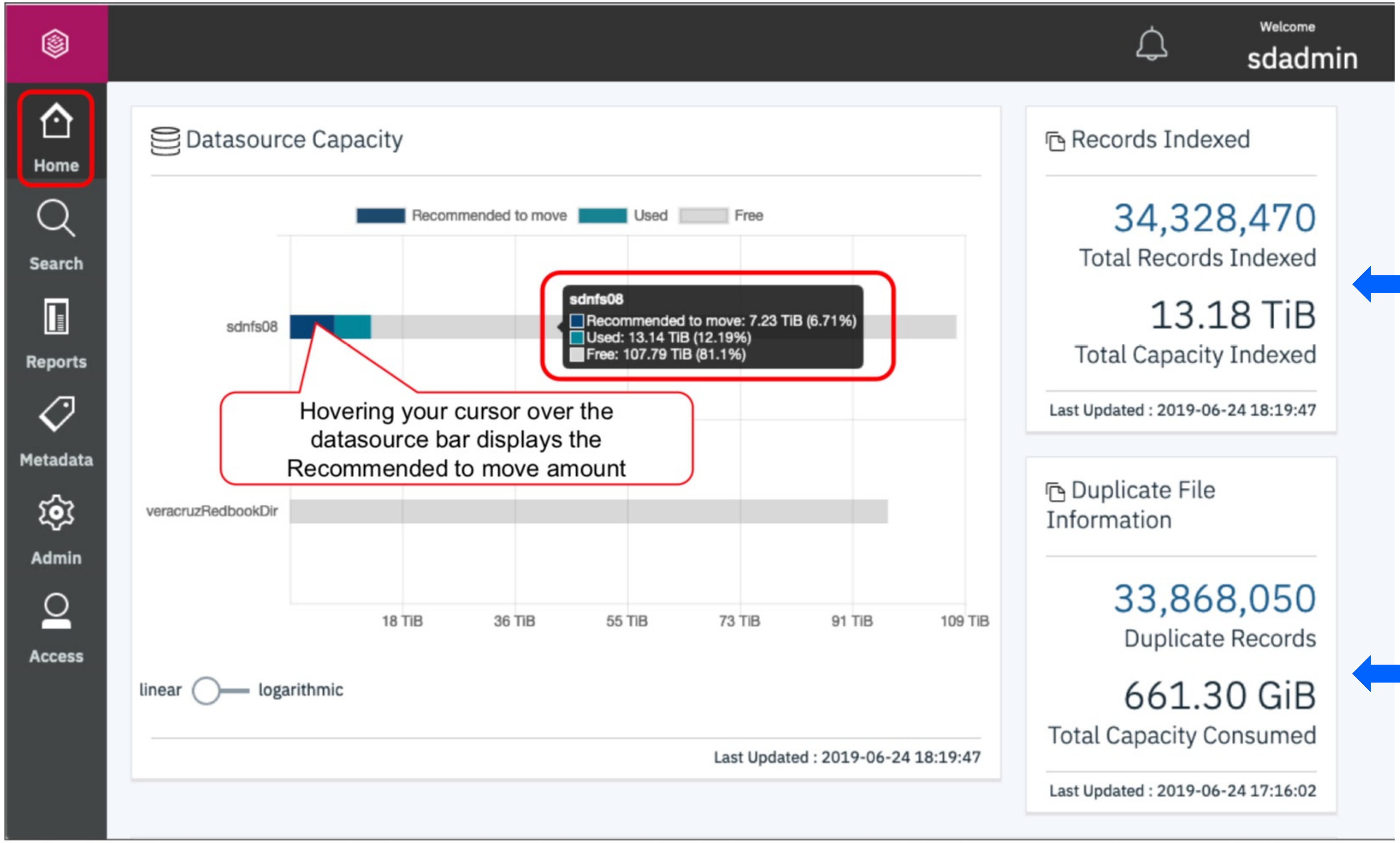
Generate reports

Customize view

Drill down

Discover in one screen duplicate records and data for archive

Data Visualization



sdnfs08
Recommended to move: 7.23 TiB (6.71%)
Used: 13.14 TiB (12.19%)
Free: 107.79 TiB (81.1%)

Hovering your cursor over the datasource bar displays the Recommended to move amount

Last Updated : 2019-06-24 18:19:47

34,328,470
Total Records Indexed

13.18 TiB
Total Capacity Indexed

Last Updated : 2019-06-24 18:19:47

33,868,050
Duplicate Records

661.30 GiB
Total Capacity Consumed

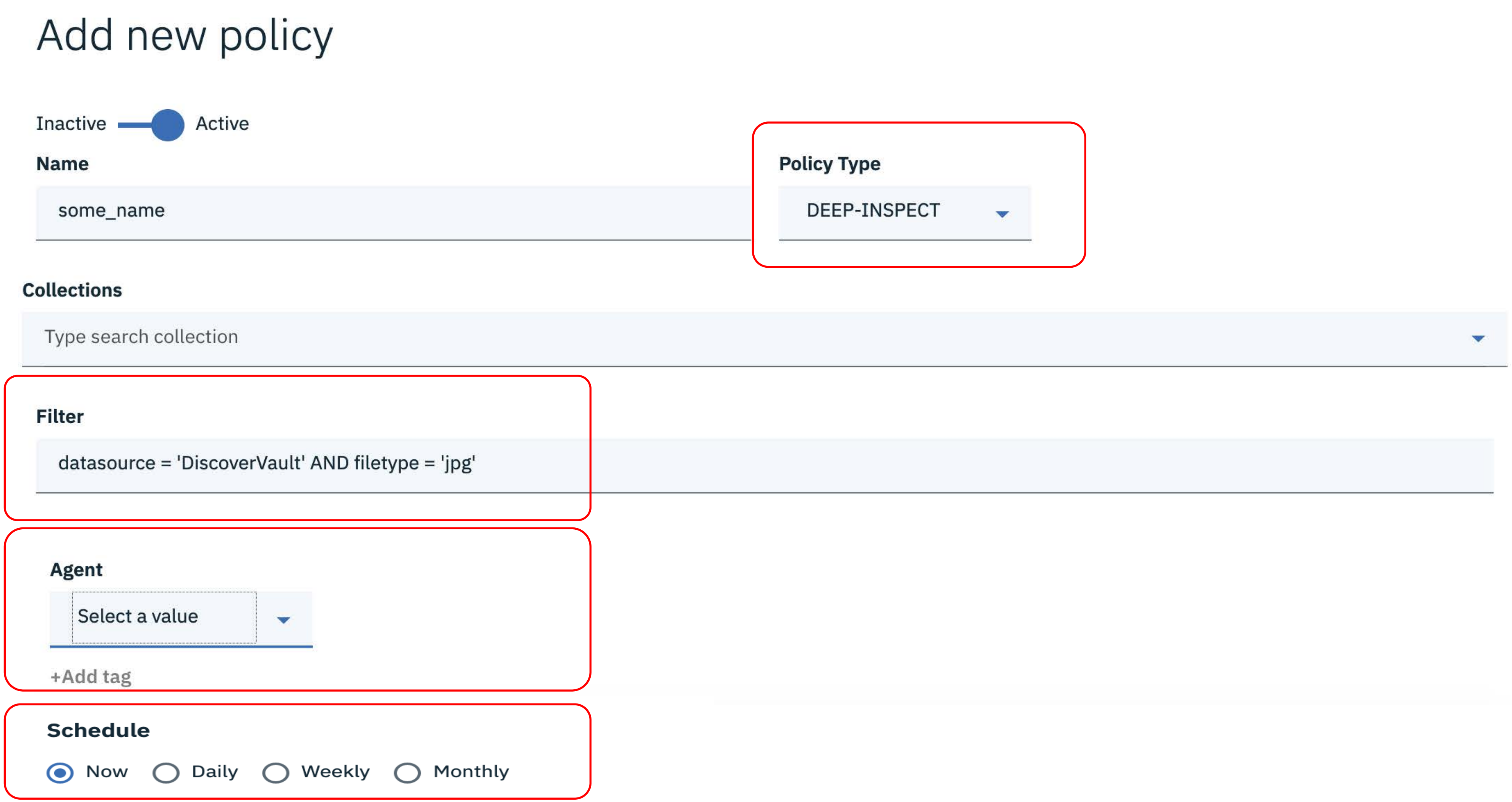
Last Updated : 2019-06-24 17:16:02

Summary of capacity

Summary of duplicate records

Create a custom “action agent” to automate a workflow

Data Activation



Add new policy

Inactive ● Active

Name

some_name

Policy Type

DEEP-INSPECT

Collections

Type search collection

Filter

datasource = 'DiscoverVault' AND filetype = 'jpg'

Agent

Select a value

+Add tag

Schedule

● Now ○ Daily ○ Weekly ○ Monthly

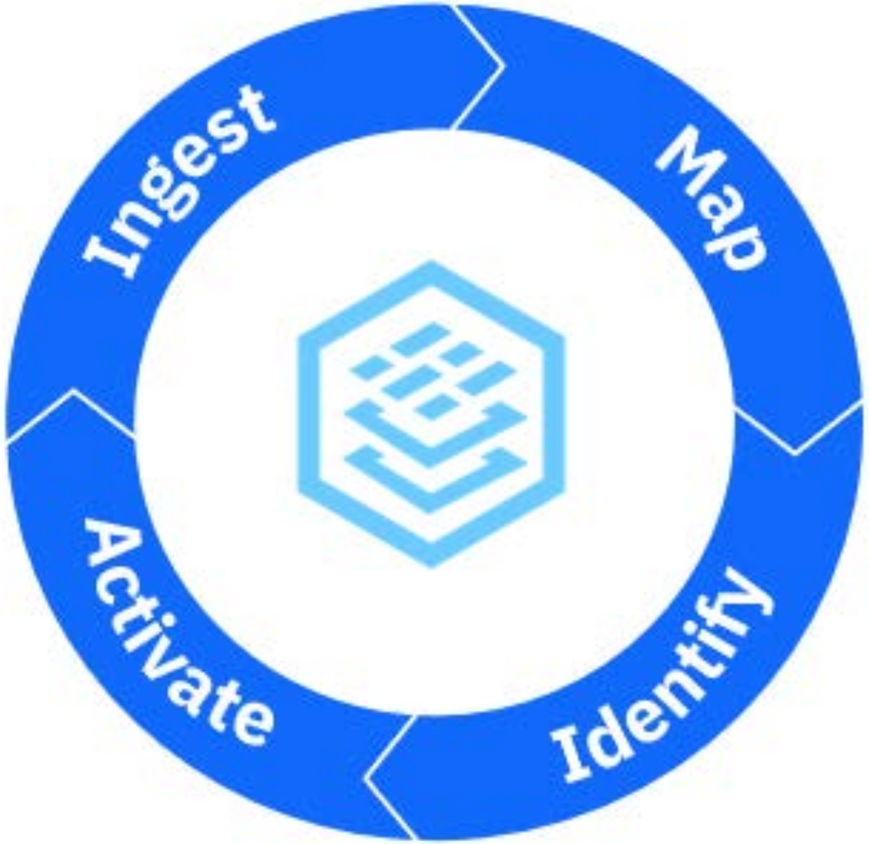
Using Spectrum Discover for metadata-fueled data analysis

Large Scale Data Ingest

- Scan billions of records per day[1]
- Scale without limits
- Multi-source support

Business-Oriented Data Mapping

- Custom data tagging
- Content based inspection
- Automatic indexing



Data Activation

- Create automated actions
- Solution blueprints
- Live event notifications
- Policy-driven workflows

Data Visualization

- Query billions of records in seconds
- Multi-faceted SQL like search
- Drilldown dashboard
- Customizable reports

[1] Up to 30K/sec; ~2.5 billion rec/day w/ IBM Spectrum Scale; ~432 million/day w/ IBM Cloud Object Storage

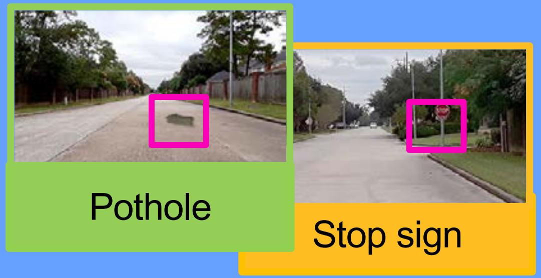
IBM Storage for AI workflows

Ingest

IBM Spectrum Scale

IBM Cloud Object Storage

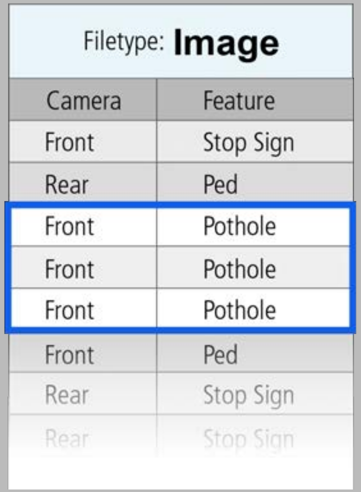
Data mapping, data visualization, data discovery



IBM Visual Insights
IBM Visual Inspector

IBM Spectrum Discover

Curate, organize, & data orchestration



Filetype: **Image**

Camera	Feature
Front	Stop Sign
Rear	Ped
Front	Pothole
Front	Pothole
Front	Pothole
Front	Ped
Rear	Stop Sign
Rear	Stop Sign

IBM Spectrum Discover

IBM Spectrum Scale

Train

NVIDIA

APACHE Spark

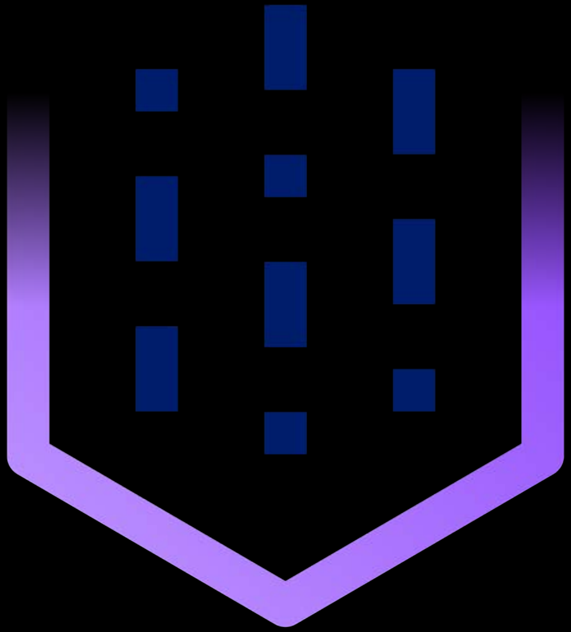
TensorFlow

CLOUDERA

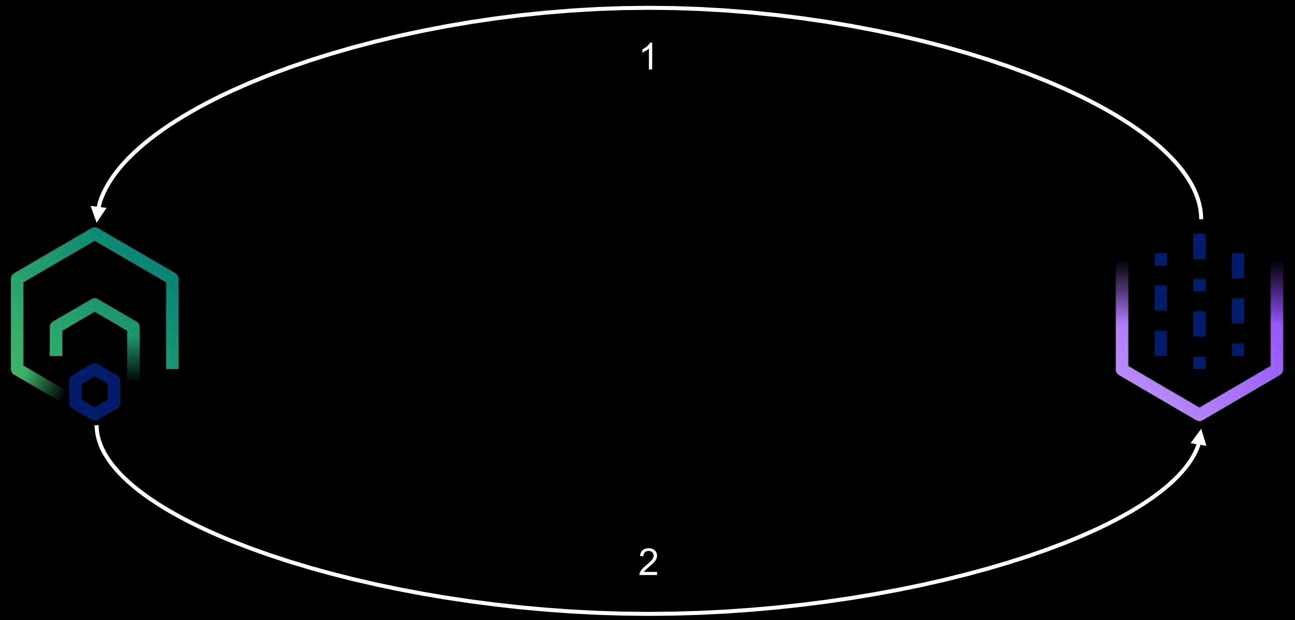
IBM Visual Insights
IBM Visual Inspector

IBM Spectrum Scale

Spectrum Discover Integration in Spectrum Scale and ESS

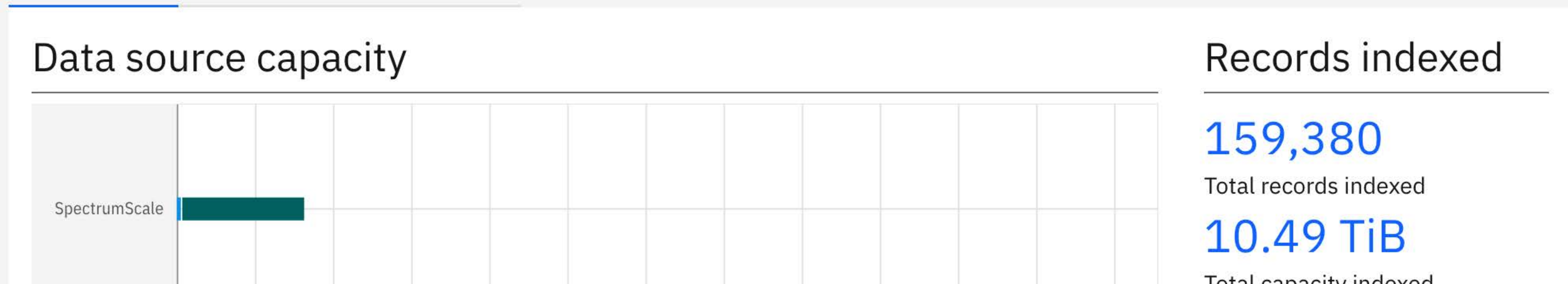


Connecting Spectrum Scale and Spectrum Discover



1. Scan Files System / Single Fileset
   - Spectrum Discover scans actively

2. Live Events
   - Spectrum Scale sends Watchfolder events to Spectrum Discover



Data source capacity

Records indexed

159,380

Total records indexed

10.49 TiB

Total capacity indexed

System metadata collected by Spectrum Discover

IBM Spectrum Scale

- Filesystem
- Site
- Platform
- Cluster
- Inode
- Owner
- Group
- uid
- gid
- Mode
- Fileset
- Path
- mtime
- atime
- ctime
- Pool
- Size
- migstatus
- migloc

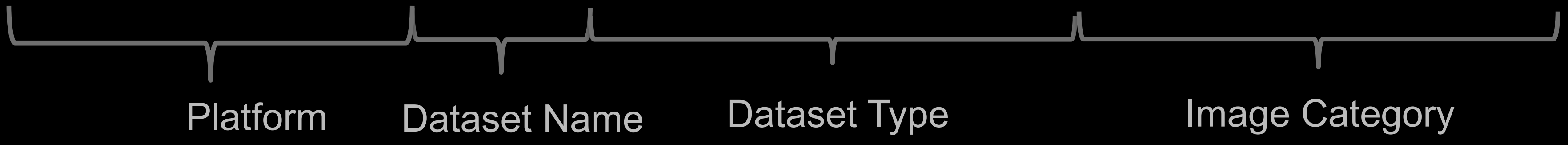
IBM Cloud Object Storage

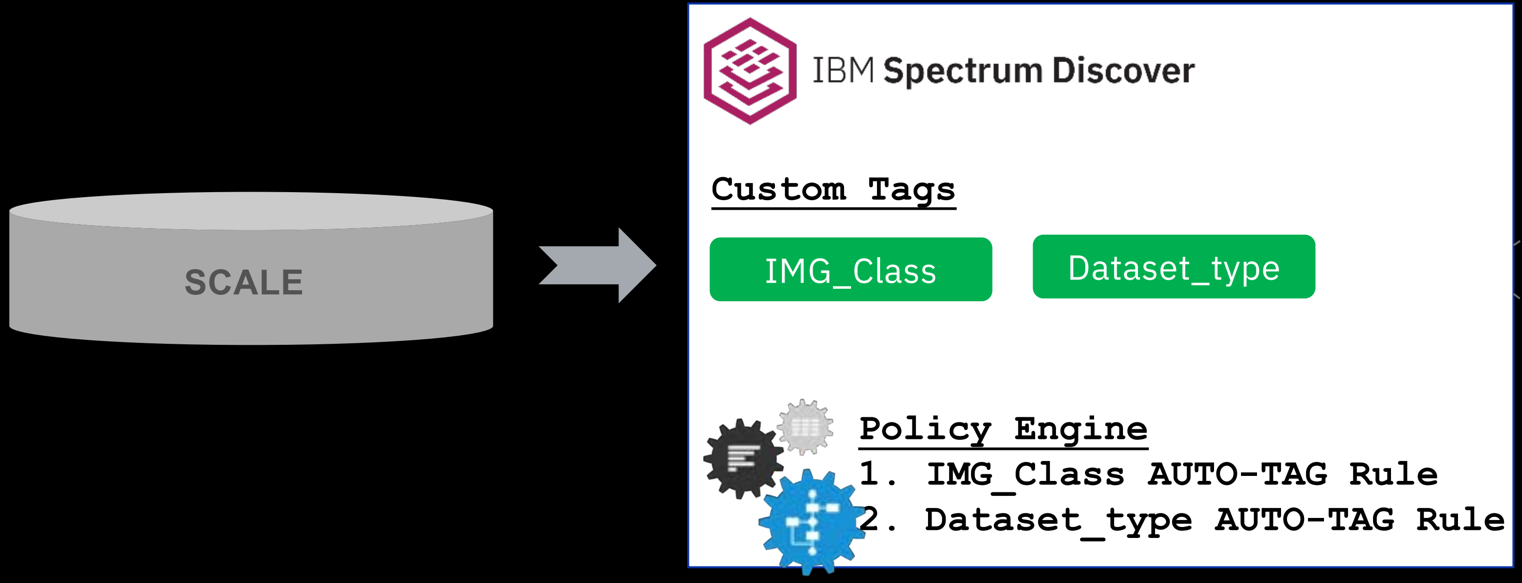
- Operation
- Bucket Name
- Object Name
- Object Length
- Object etag
- Content Type
- Bucket UUID
- System UUID

Leverage filesystem patterns to automatically create custom tags for dataset management

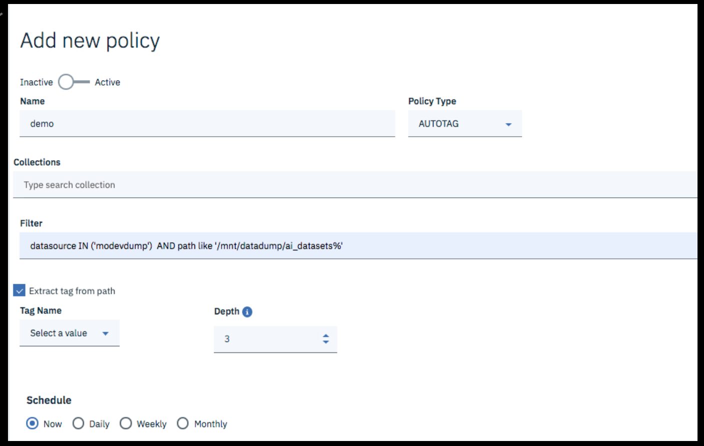
- Automatically parse file system sub-directories and insert as custom tags into Spectrum Discover
- Track and manage data across storage pools by custom tags

```
/mnt/datadump/ai_datasets/POWER-AI-VISION/EPRI/Training Set Final/Conductor Damaged
```





1. IMG_Class AUTO-TAG Rule
2. Dataset_type AUTO-TAG Rule



Add new policy

Inactive ○— Active

Name: demo

Policy Type: AUTOTAG

Collections: Type search collection

Filter: datasource IN ('modevdump') AND path like '/mnt/datadump/ai_datasets%'

☑ Extract tag from path

Tag Name: Select a value

Depth: 3

Schedule

● Now ○ Daily ○ Weekly ○ Monthly

Spectrum Scale ILM Policies

Files written in file system are placed storage pools

- Storage poola are a collection of disks of the same type

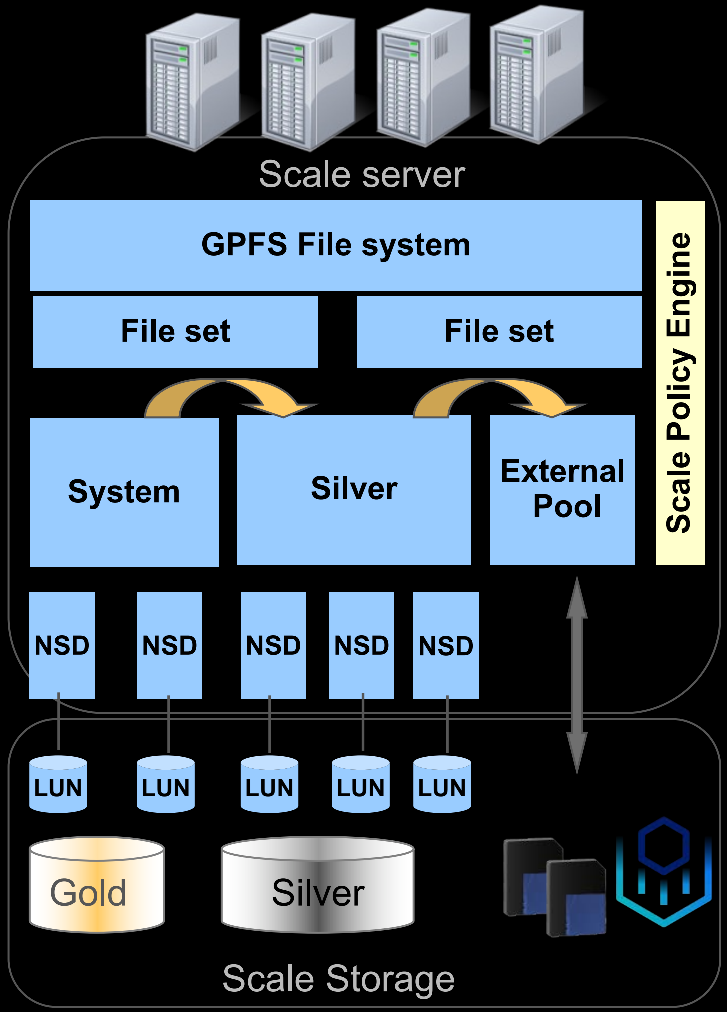
Placement policy controls where files are placed

- Applied during file creation

Migration policies control transparent migration of files from one pool to another

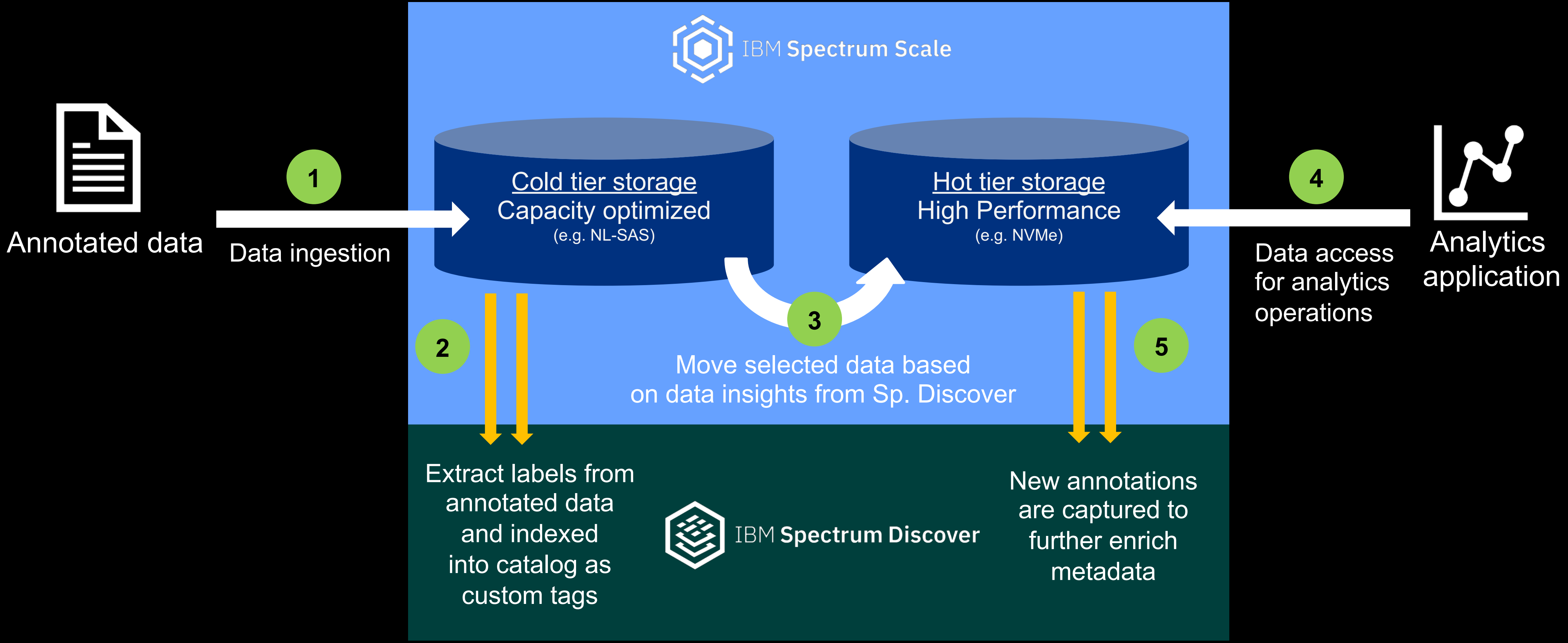
- Applied during life cycle
- Migrated files can be transparently accessed
- Files can also be migrated to external pools such as
  - Tape
  - Object Storage

Policy engine applies and executes policies based on metadata or manually.



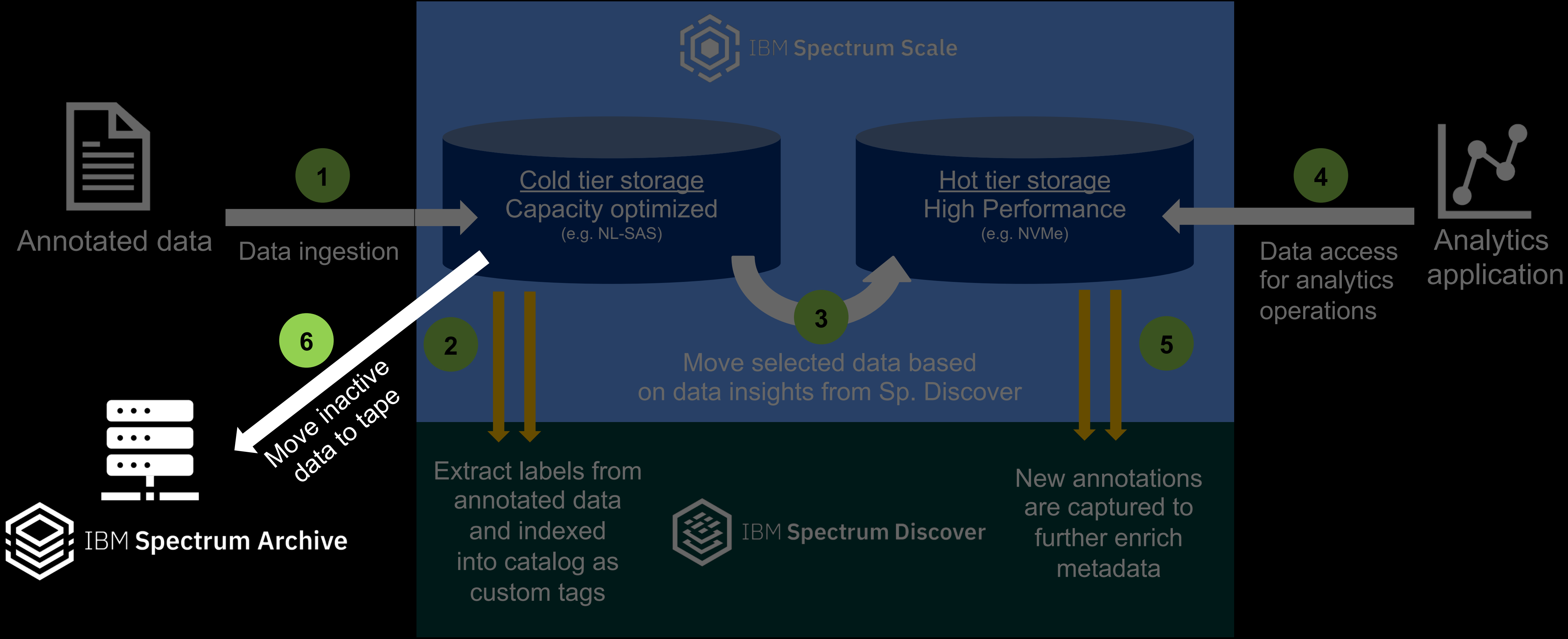
Data placement optimization for Spectrum Scale and ESS

- Leverage the insights from Spectrum Discover to drive better data management in Spectrum Scale
- Goes beyond the system metadata and extended attributes used by the current Spectrum Scale policy engine.


IBM Spectrum Scale
Annotated data
1
Data ingestion
Cold tier storage
Capacity optimized
(e.g. NL-SAS)
Hot tier storage
High Performance
(e.g. NVMe)
4
Data access for analytics operations
Analytics application
3
2
5
Move selected data based on data insights from Sp. Discover
Extract labels from annotated data and indexed into catalog as custom tags
IBM Spectrum Discover
New annotations are captured to further enrich metadata

Further data placement optimization for Spectrum Archive Enterprise Edition

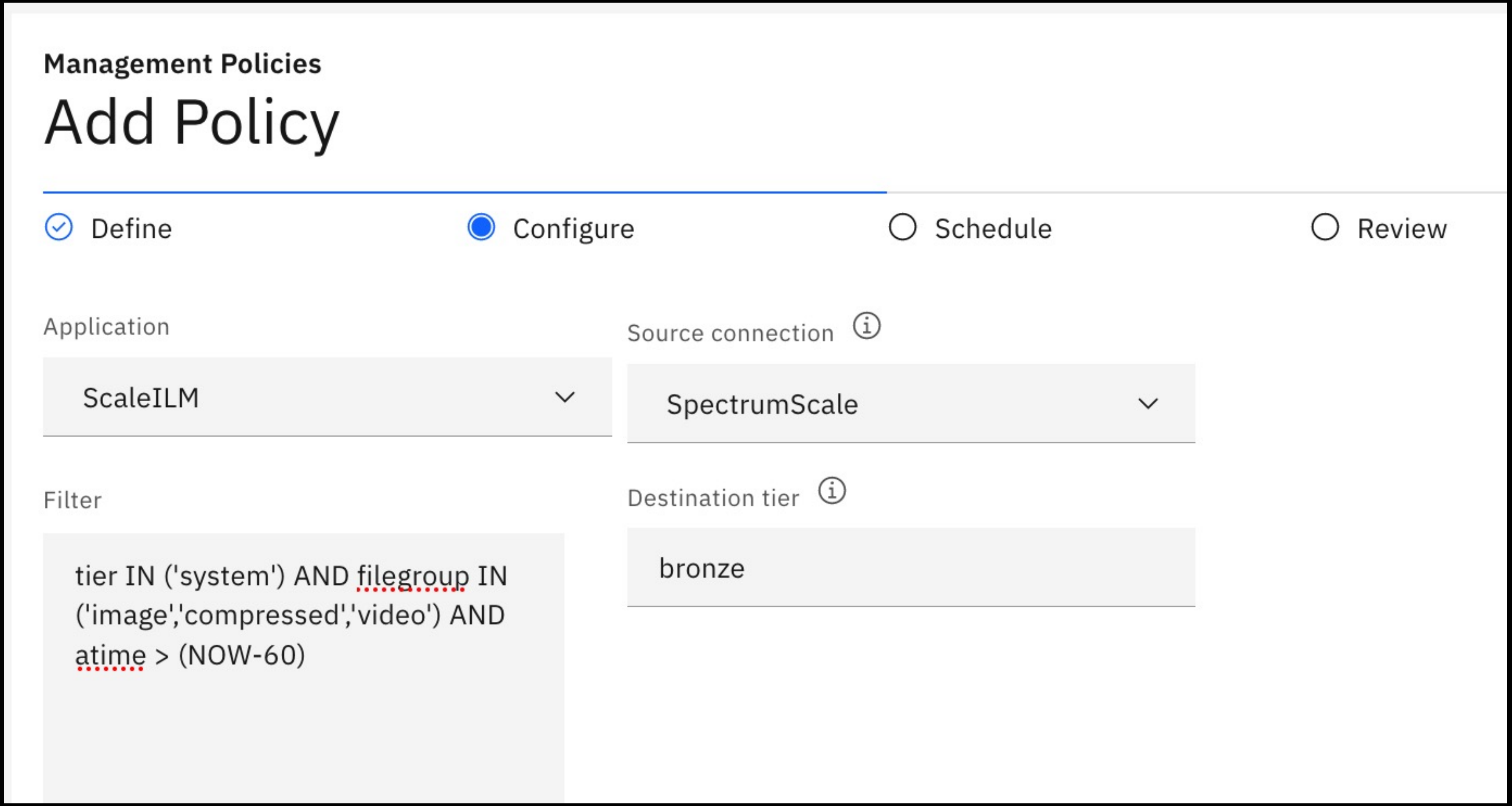
- Move data to/from Spectrum Archive Tape
- A searchable catalog is maintained that includes the tape


IBM Spectrum Scale
Annotated data
1
Data ingestion
Cold tier storage
Capacity optimized
(e.g. NL-SAS)
Hot tier storage
High Performance
(e.g. NVMe)
4
Data access for analytics operations
Analytics application
3
Move selected data based on data insights from Sp. Discover
2
5
6
Move inactive data to tape
Extract labels from annotated data and indexed into catalog as custom tags
IBM Spectrum Discover
New annotations are captured to further enrich metadata
IBM Spectrum Archive

Trigger Scale Data tiering

Spectrum Discover triggers ILM Policies

- Tier data to different storage pool based on system metadata



Management Policies

Add Policy

Application: ScaleILM

Source connection: SpectrumScale

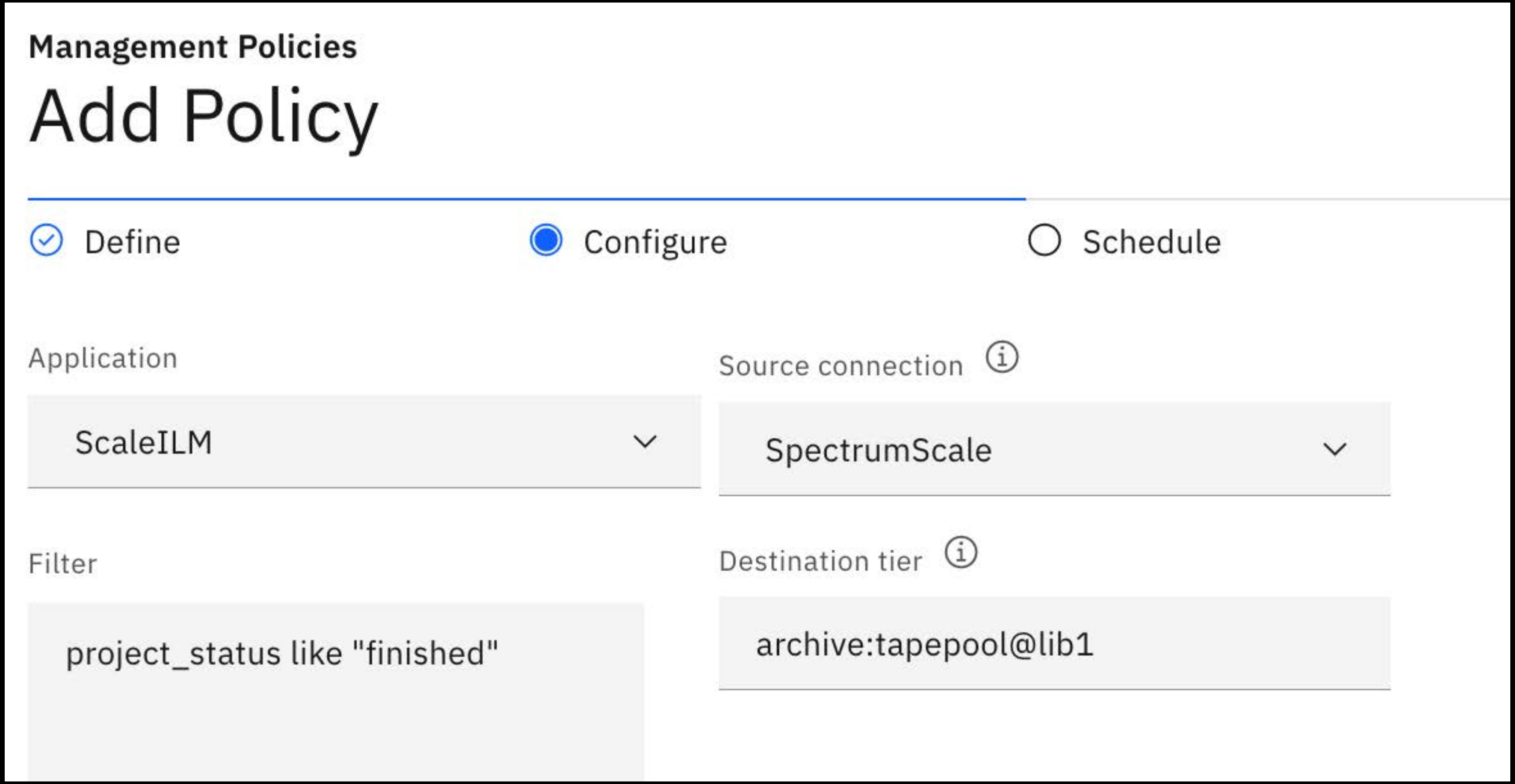
Filter: tier IN ('system') AND filegroup IN ('image','compressed','video') AND atime > (NOW-60)

Destination tier: bronze

Trigger Scale Data tiering

Spectrum Discover triggers ILM Policies

- Tier data to different storage pool based on system metadata
- Tier data to different storage pool based on custom metadata



Management Policies

Add Policy

Application: ScaleILM

Source connection: SpectrumScale

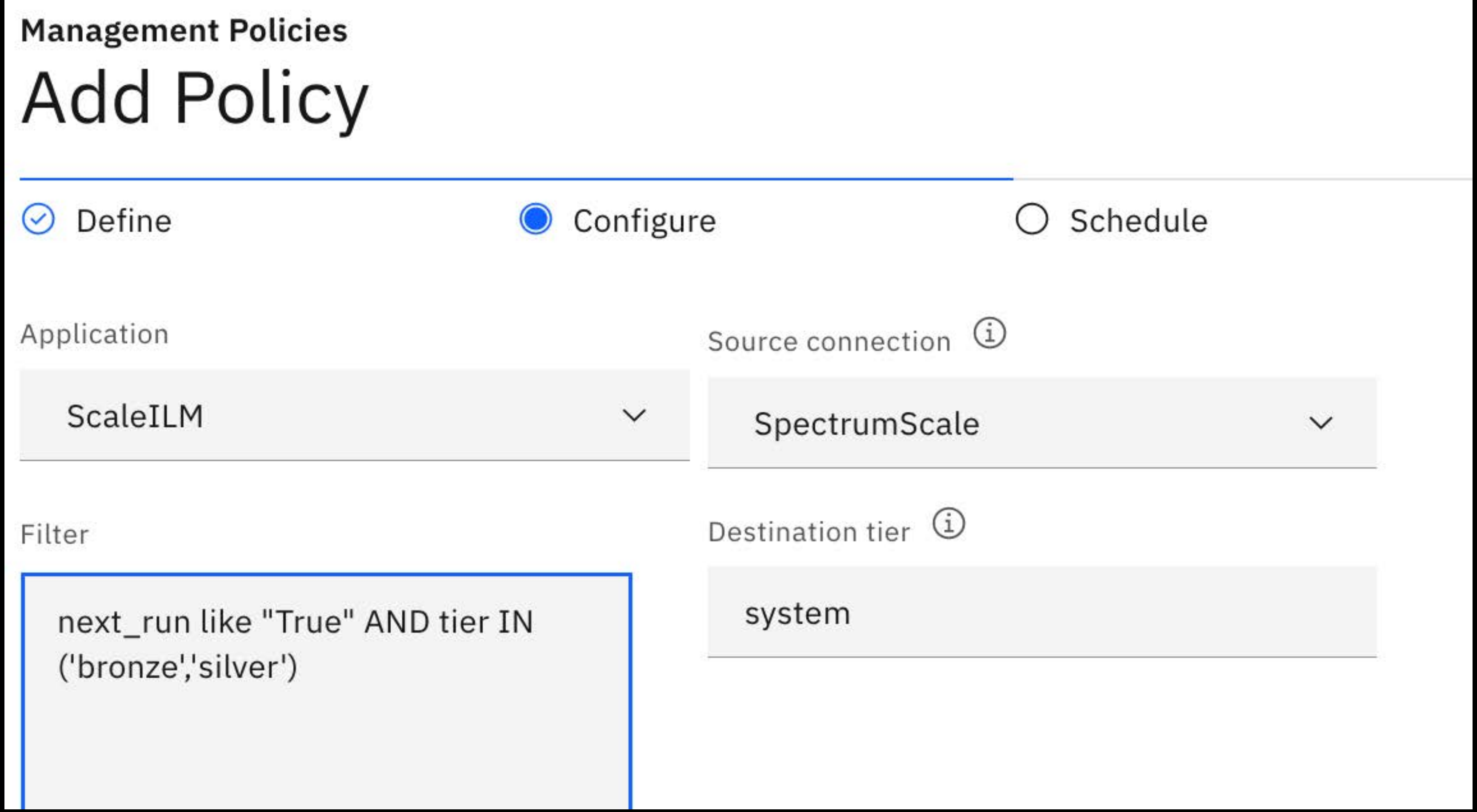
Filter: project_status like "finished"

Destination tier: archive:tapepool@lib1

Trigger Scale Data tiering

Spectrum Discover triggers ILM Policies

- Tier data to different storage pool based on system metadata
- Tier data to different storage pool based on custom metadata
- Recall data to hot tier based on custom metadata



Management Policies

Add Policy

✓ Define ● Configure ○ Schedule

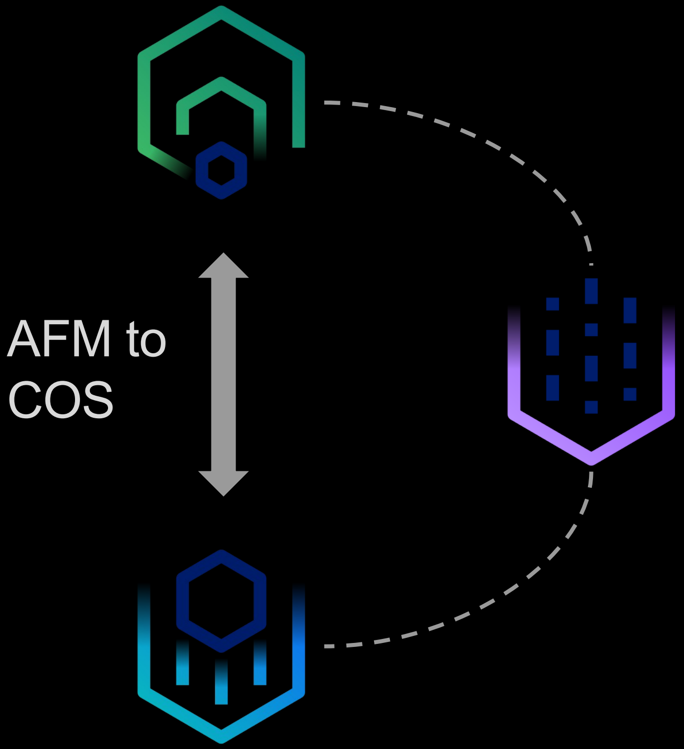
Application: ScaleILM

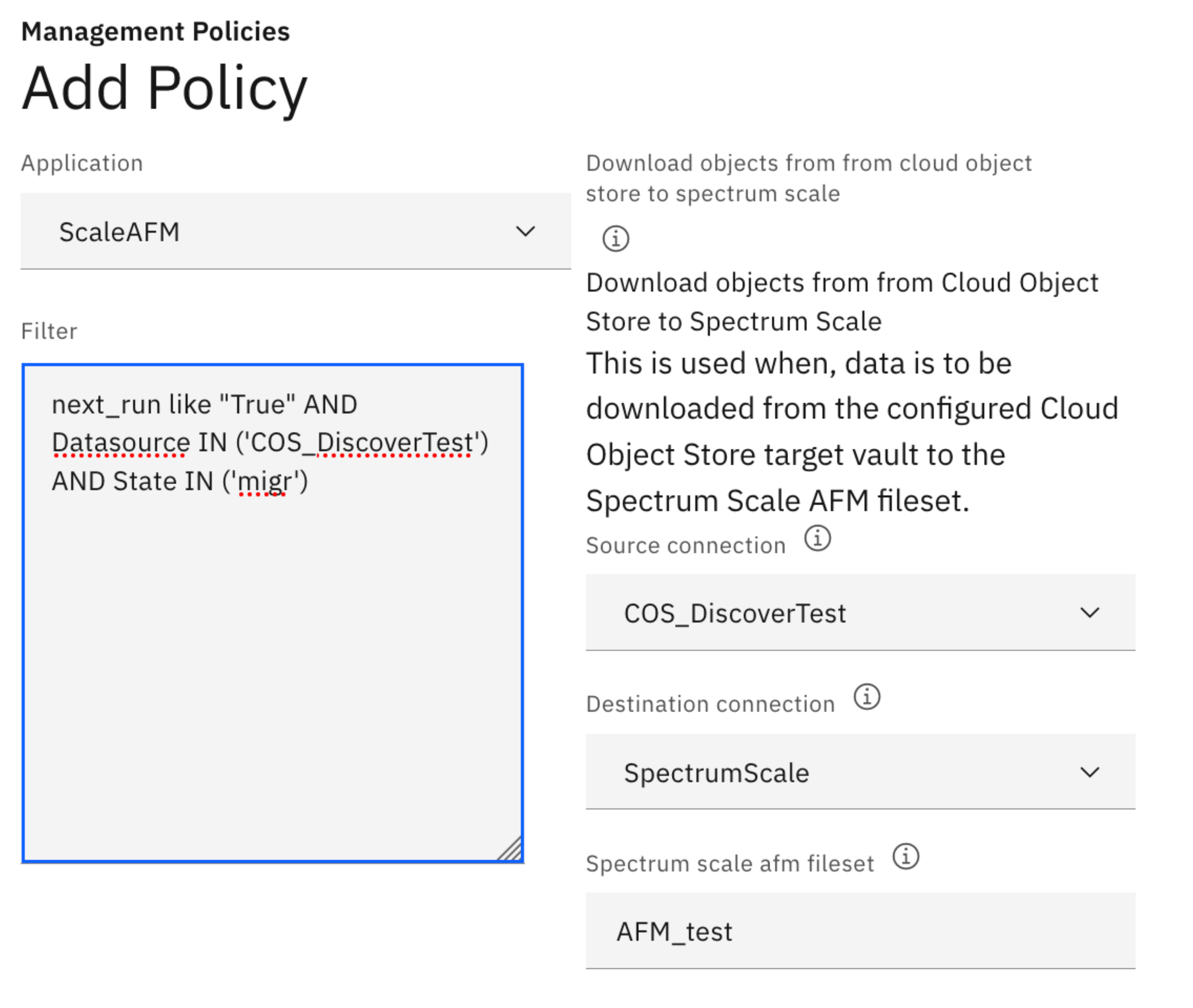
Source connection: SpectrumScale

Filter: next_run like "True" AND tier IN ('bronze','silver')

Destination tier: system

Discover and Scale AFM





Add Policy

Application

ScaleAFM

Filter

```
next_run like "True" AND
Datasource IN ('COS_DiscoverTest')
AND State IN ('migr')
```

Download objects from from cloud object store to spectrum scale

Download objects from from Cloud Object Store to Spectrum Scale
This is used when, data is to be downloaded from the configured Cloud Object Store target vault to the Spectrum Scale AFM fileset.

Source connection: COS_DiscoverTest

Destination connection: SpectrumScale

Spectrum scale afm fileset: AFM_test

Use Case: Tumor Classification

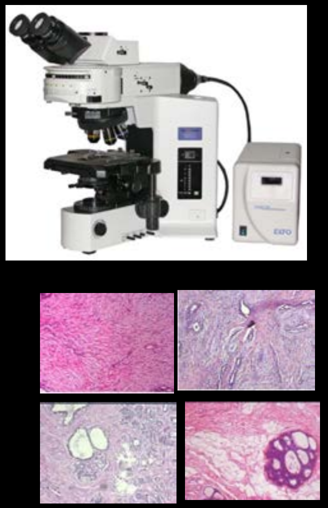
Event driven architecture to automatically classify and catalog biopsies of breast cancer tumors using PowerAI Vision inference model, Spectrum Discover, and Spectrum Scale / ESS

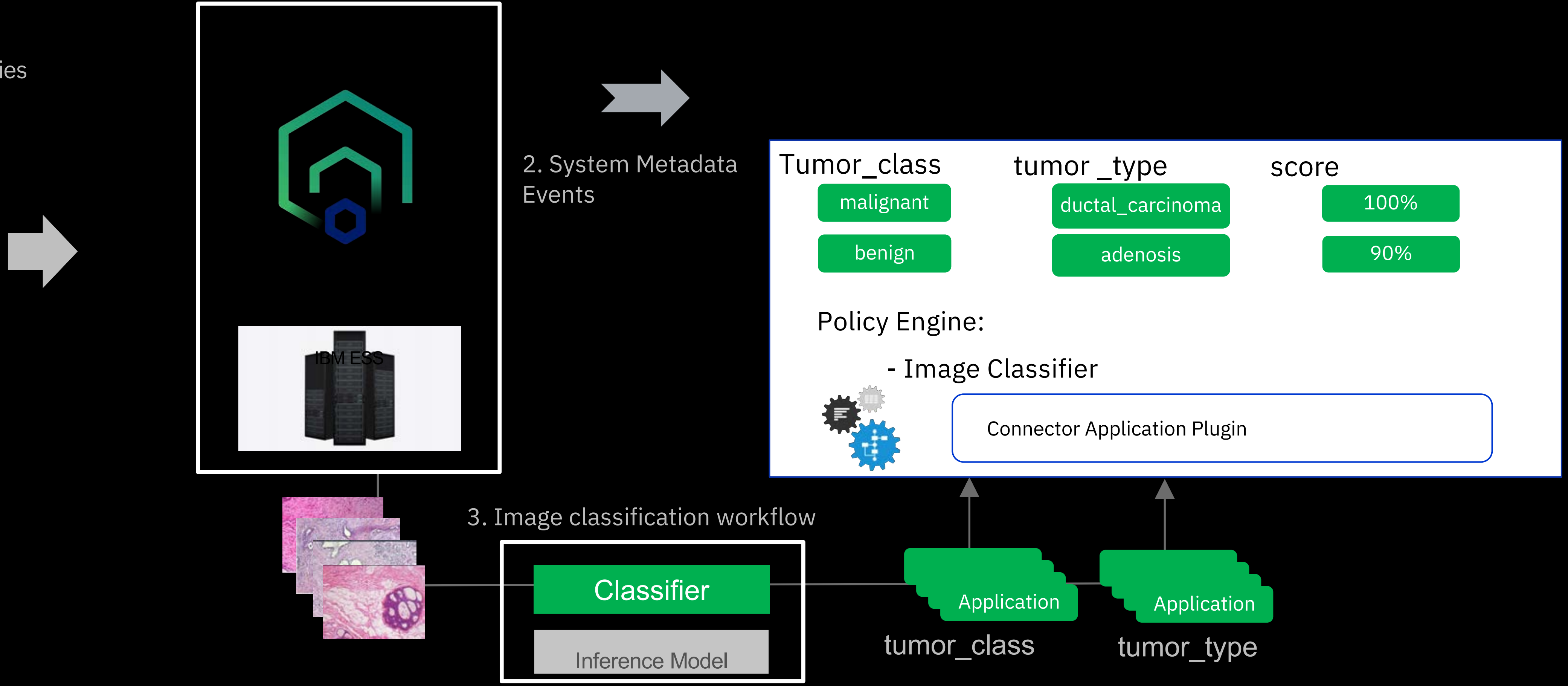
1. New imaging data ingested into Spectrum Scale / ESS

2. Storage sends Spectrum Discover system metadata events when new imaging data is ingested and Spectrum Discover builds catalog

3. Spectrum Discover policy automatically reads new imaging data from source storage, passes to the image classification model, captures results and indexes into Spectrum Discover

Breast Cancer Tumor Biopsies





Tumor_class	tumor _type	score
malignant	ductal_carcinoma	100%
benign	adenosis	90%

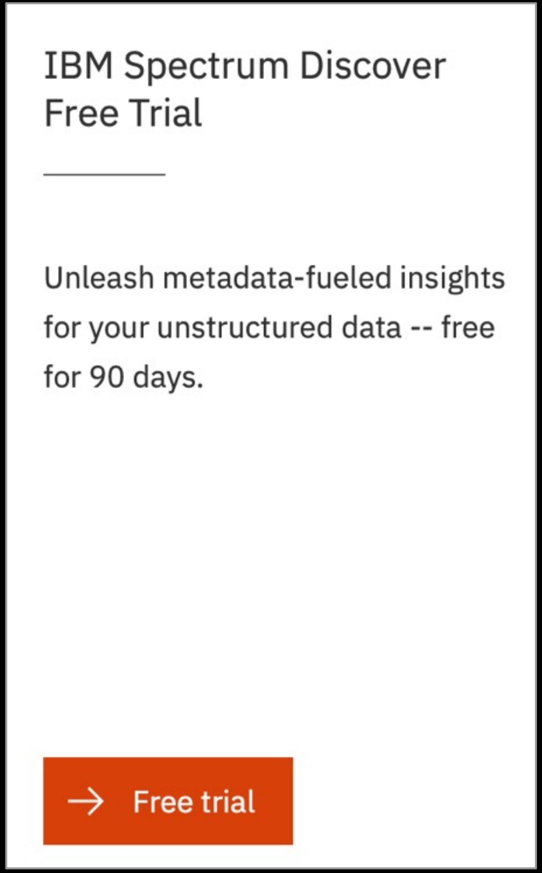
Policy Engine:

- Image Classifier

Connector Application Plugin

Free 90-day trial

Experience for yourself the game-changing insights possible with IBM Spectrum Discover

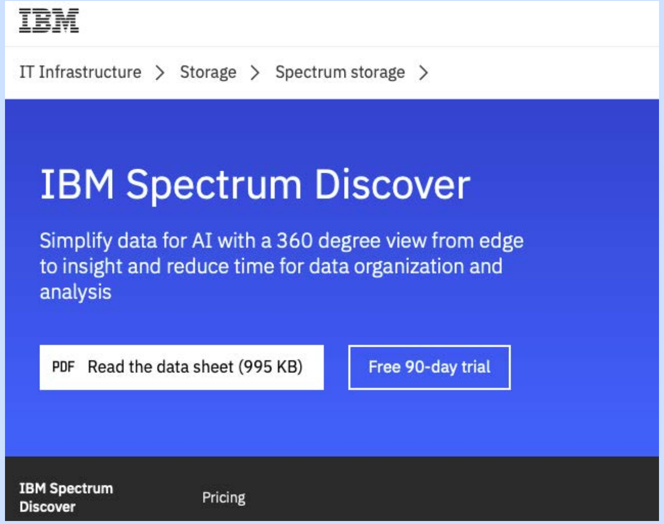


IBM Spectrum Discover Free Trial

Unleash metadata-fueled insights for your unstructured data -- free for 90 days.

www.ibm.com/marketplace/spectrum-discover

Learn more about Spectrum Discover



IBM

IT Infrastructure > Storage > Spectrum storage >

IBM Spectrum Discover

Simplify data for AI with a 360 degree view from edge to insight and reduce time for data organization and analysis

PDF Read the data sheet (995 KB)

Free 90-day trial

IBM Spectrum Discover

Pricing

Web Page and Resources

https://www.ibm.com/products/spectrum-discover

Redbooks

Cataloging Unstructured Data in IBM Watson Knowledge Catalog with IBM Spectrum Discover

Cloud

Analytics

Storage

IBM

Redpaper

http://www.redbooks.ibm.com/abstracts/redp5603.html?Open

Redbooks

IBM Spectrum Discover

Metadata Management for Deep Insight of Unstructured Storage

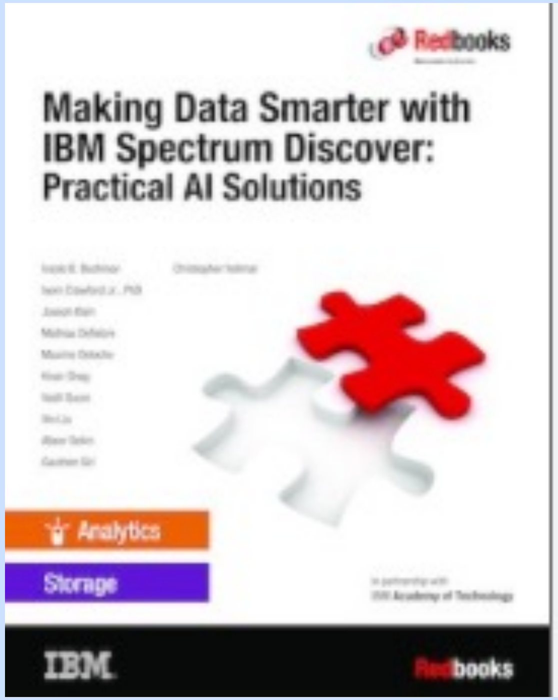
Analytics

Storage

IBM

Redpaper

http://www.redbooks.ibm.com/redpapers/pdfs/redp5550.pdf



Redbooks

Making Data Smarter with IBM Spectrum Discover: Practical AI Solutions

Analytics

Storage

IBM

Redbooks

http://www.redbooks.ibm.com/abstracts/sg248488.html?Open

Thank you for using
IBM Spectrum Scale with Spectrum Discover!