

# IBM Spectrum Discover

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

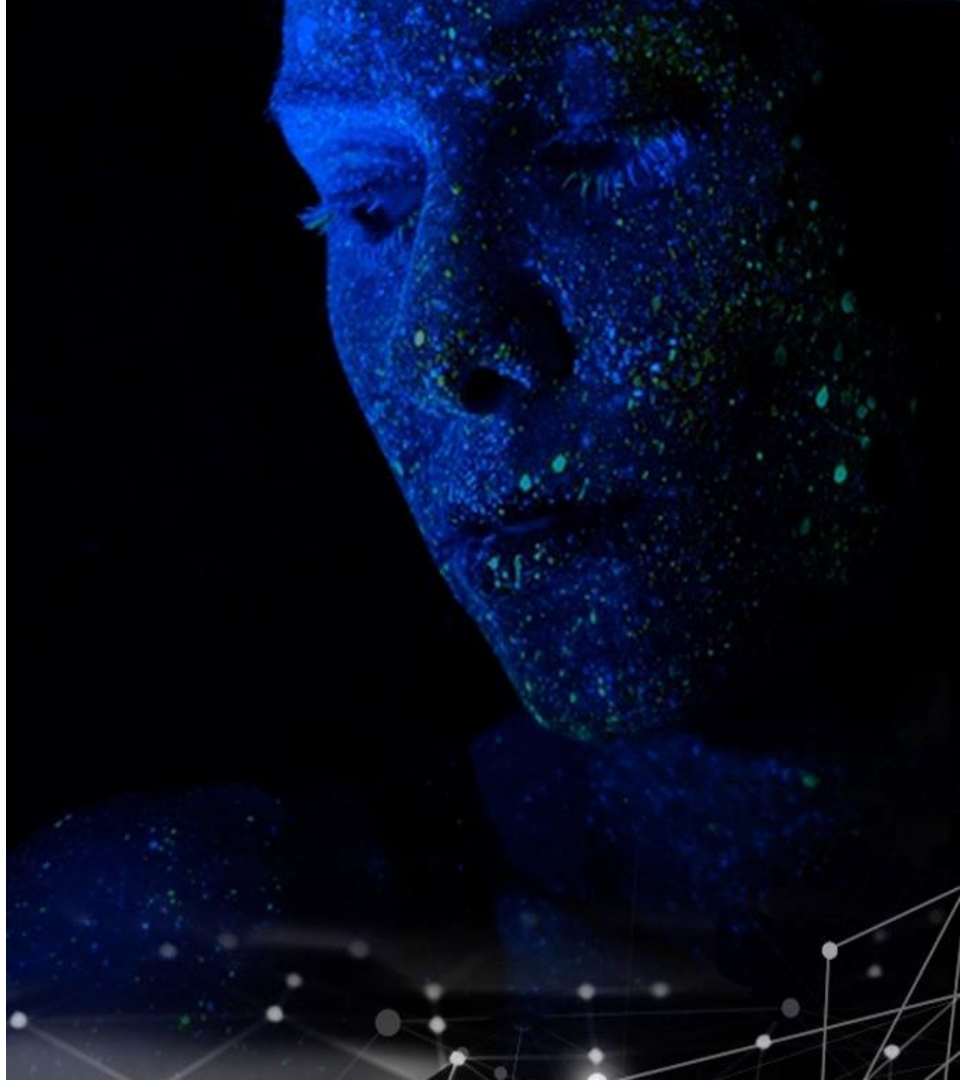
Indulis Bernsteins  
Consulting Systems Architect  
IBM UK





# Agenda

- Problem Statement
- Spectrum Discover Overview
- Use Case Discussion
- Demo

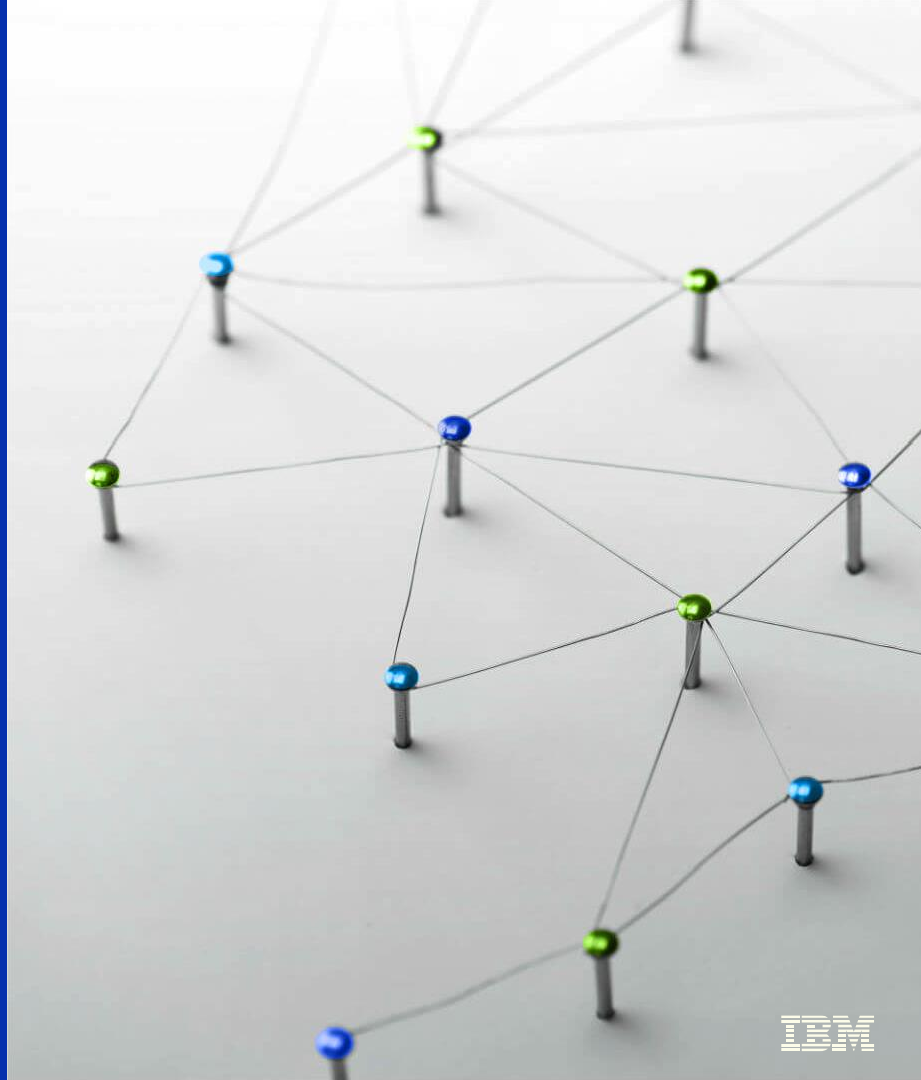




# Unstructured Data is Hard to Manage

For 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

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)

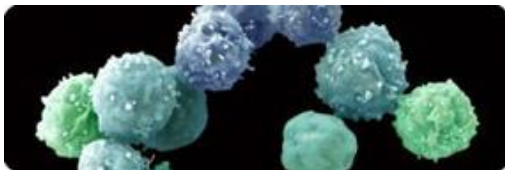
Custom metadata tags

AI derived metadata

System Metadata

- Location
- Size
- Owner
- Group
- Permissions
- Last-Modified
- ...

Biomedical

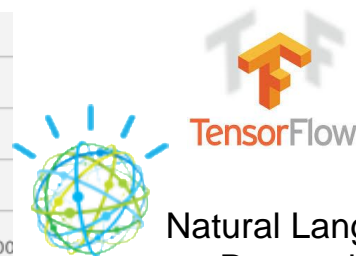


Image

	File Size
	1.1 MB
	Dimensions
	1280 x 1024 pixels
	File Date
	Aug 22, 2011, 9:42 AM
	JPEG Quality
	96 (444)
Unique ID	
31d24e7a2fe0190600000000000000	
Software	
Adobe Photoshop CS5 Macintosh	

IBM PowerAI Vision

PYTORCH



Natural Language  
Processing

figure  
eight

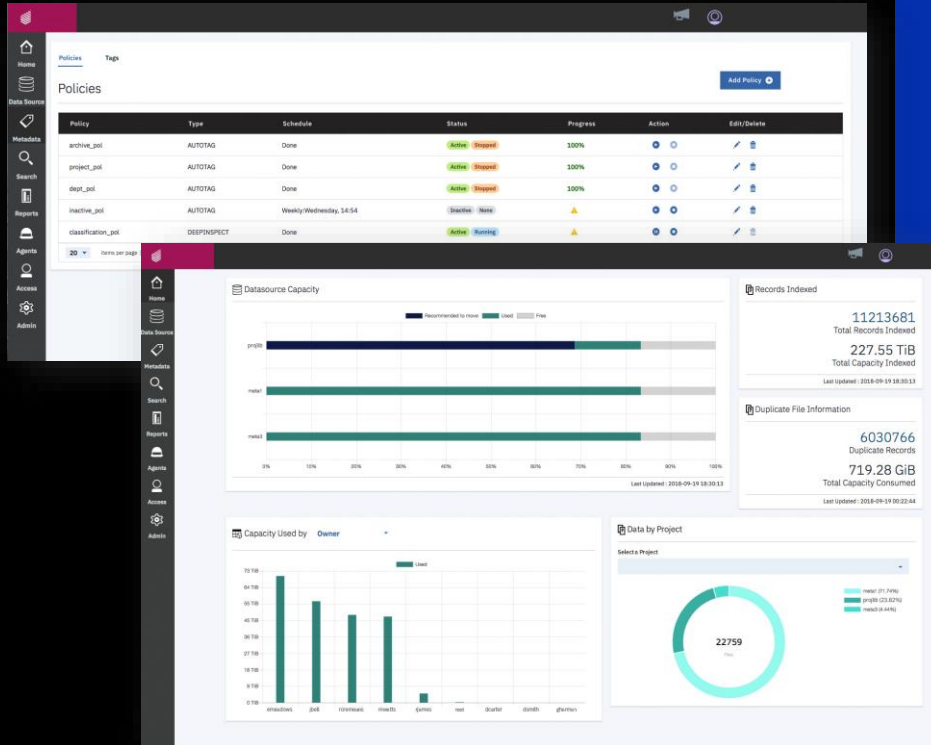




# IBM Spectrum Discover

## 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 Accelerates Value of data

## For Optimization

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

*Improve storage utilization*

## For Governance

- Ensure data is consistent with governance policies
- Reduce risk buried in unstructured data stores
- Speed investigations for legal discovery & regulatory audits

*Mitigate risk & improve data quality*

## For Analytics

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

*Reduce time to accuracy & results*



# Multiple concurrent ways to leverage Spectrum Discover

## Large-scale Analytics/AI/ML

- Data mapping
- Data discovery
- Dataset identification
- Data pipeline progression

## Data Optimization

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

## Data Governance

- Data inspection and classification
- Label sensitive data for compliance
- Data clean-up

## Data Management

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

IT admin / architect

Application user / data admin

Data scientist



# IBM Spectrum Discover Environment

## File, Object, Backup, and Archive Storage



IBM  
Spectrum  
Scale



IBM  
Spectrum  
Protect



IBM Cloud  
Object  
Storage



IBM  
Spectrum  
Archive



S3



## Data Insight



IBM Spectrum Discover



Search



Reporting



Dashboard

- Simple to deploy (VMware virtual appliance)
- Metadata curation
- Custom metadata tagging
- Automatic indexing
- Policy-Engine
- Action Agent API

## Activation & Optimization

### Large-Scale Analytics

- Data discovery
- Dataset identification
- Data pipeline progression

### Data Governance

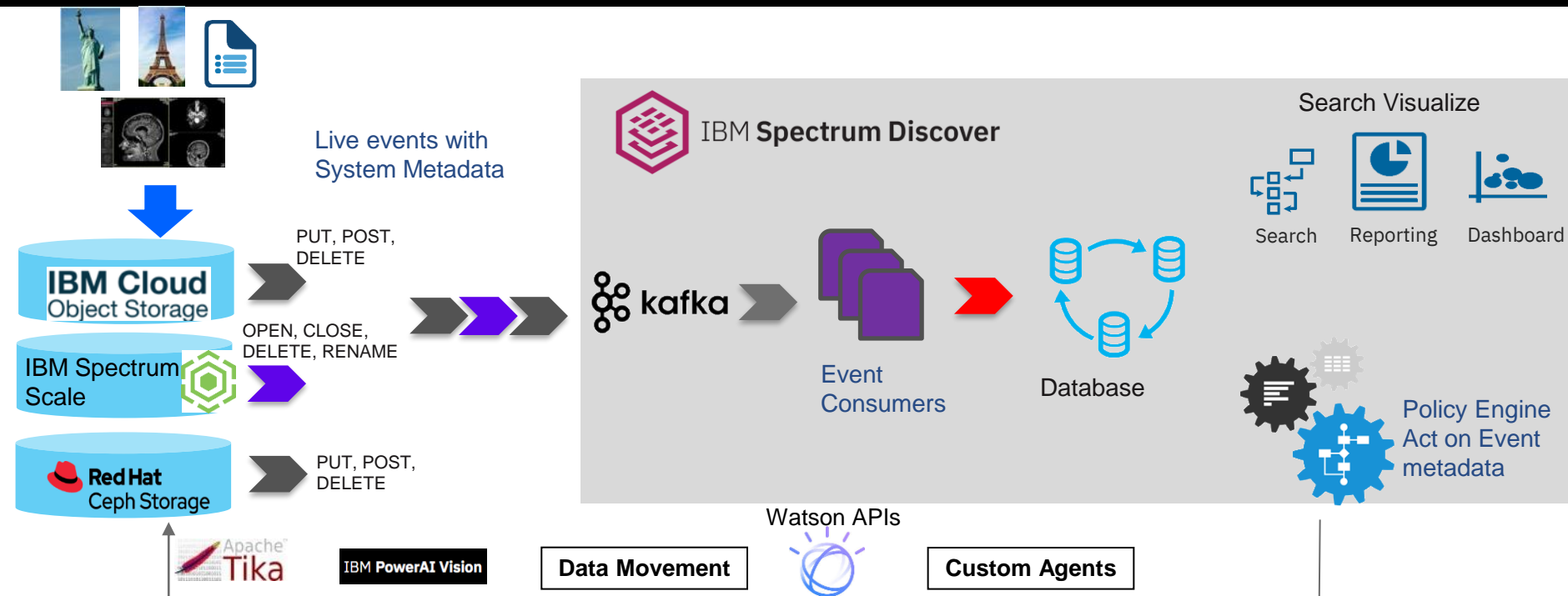
- Data inspection
- Data classification
- Data clean-up

### Data Optimization

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



# Metadata event driven architecture



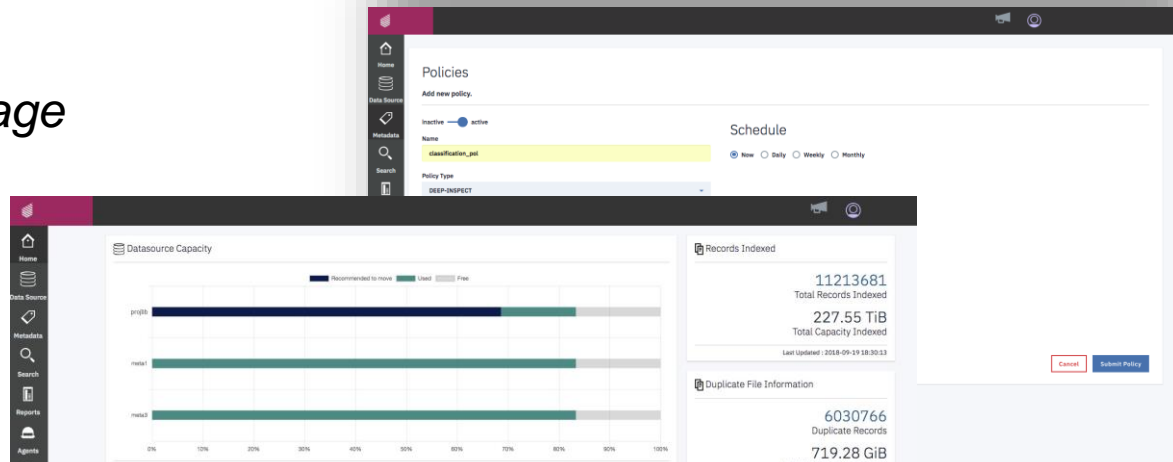
## Harvest events from heterogeneous sources

- Transparent – events generated by underlying storage platform No modification to applications or user behavior
- Real time, low overhead – immediate visibility of user actions Rapid response from analytics
- Allows extreme storage scale – events proportional to user activity, not to total size of data being monitored
- Take action on event metadata



# Spectrum Discover enables metadata management for an AI infrastructure

*Unified metadata and data insights for file and object storage on-premises and in the cloud*



## Discover

Automatically ingest and index system metadata from heterogeneous file and object storage systems on-prem and in the cloud

## Classify

Automatically identify and classify data, including sensitive and personally identifiable information

## Label

Enrich data with system and custom metadata tags that increase the value of that data

## Find

Find data quickly and easily by searching catalogs of system and custom metadata



# Storage Optimization with Spectrum Discover



## Optimization – Improve Storage Utilization

### Key questions...

- How is my data aging?
- What type of data do I have?
- What is the size distribution of my data?
- Do I have duplicate data in my environment?
- How can I map this data to my business constructs?

Leverage Spectrum Discover built in analytics to identify ROT data

1. Ingest system metadata for files and objects
2. Leverage default analytics and generate reports
3. Customize analytics and reports
4. Map analytics against one or more system metadata attributes



# Proven value from PoCs

## Insights



254

million files  
across 3 filesystems  
and 49 projects  
identified & tagged

## Savings



45%

data identified  
as inactive &  
candidate for archive

## Optimization



84

users with 100% of data  
inactive & identified  
for archive or  
backup/delete

Actual results from PoC conducted with beta client in heterogeneous environment –  
a major public health institution doing genomics research



# File and Object Size Distribution Analytics

- Leverage size bucketing , visualization, and drill down search
- Default Bucketing

Extra Small	Small	Medium	Large	Extra Large
<4KiB	4KiB -1MiB	1MiB – 1GiB	1GiB – 1TiB	>1TiB

## Customize Bucket Ranges

### Modify Bucket

Please make sure that the maximum value for each bucket is greater than the value assigned to the previous bucket

**Bucket Name**

SizeRange

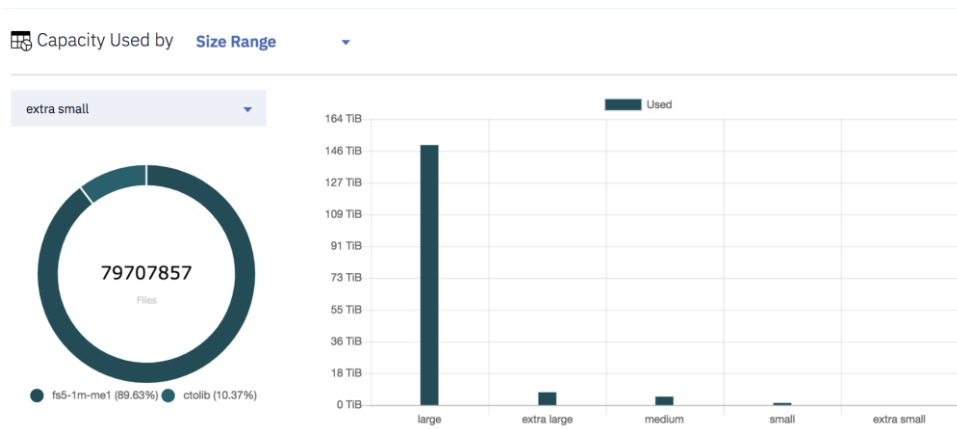
☒ extra small

Values less than

☒ small

Between previous value

## Visualize Capacity Usage by size bucketing





# File and Object Size Distribution Analytics

## Size range drilldown search

←

View results by:

Results:

	sizerange	Total Files	Total Size
<input type="checkbox"/>	extra large	4	7.32 TiB
<input type="checkbox"/>	large	8,666	148.88 TiB
<input type="checkbox"/>	small	73,124,914	1.22 TiB
<input type="checkbox"/>	medium	189,727	4.82 TiB
<input type="checkbox"/>	extra small	79,707,857	4.15 GiB

Items per page: 20 | 1-5 of 5 items

1 of 1 pages < 1 >

### SIZE RANGE

- ☐ extra large ( 4 )
- ☐ medium ( 189,727 )
- ☐ small ( 73,124,914 )
- ☐ large ( 8,666 )
- ☐ extra small ( 79,707,857 )

### TIMESINCEACCESS

Convert to individual records

	path	filename	datasource	owner	fileset	size
<input type="checkbox"/>	/ctolib/ceccleston/	HG00419.mapped.ILLUMINA.bwa.CHS.low_coverage.20130415.bam	ctolib	coswald	root	34987828313.000
<input type="checkbox"/>	/ctolib/ceccleston/	HG00557.mapped.ILLUMINA.bwa.CHS.low_coverage.20130415.bam	ctolib	coswald	root	34764596480.000
<input type="checkbox"/>	/ctolib/pmcgann/Invincible/sv_discovery_indexes/smrtp/CHS/ftp.sra.ebi.ac.uk/vol1/ERA562/ERA562I05/pacbio_hdf5/	m150823_120604_42216_c100828382550000001823180911251523_s1_p0.2.bax.h5	ctolib	jharkness	pmcgann	4629168610.000
<input type="checkbox"/>	/ctolib/ceccleston/	HG00288.mapped.ILLUMINA.bwa.FIN.low_coverage.20130502.bam	ctolib	coswald	root	27251464467.000
<input type="checkbox"/>	/ctolib/pmcgann/Invincible/sv_discovery_indexes/smrtp/CHS/ftp.sra.ebi.ac.uk/vol1/ERA562/ERA562I05/pacbio_hdf5/	m150902_035357_42220_c100828042550000001823175811251500_s1_p0.3.bax.h5	ctolib	jharkness	pmcgann	3615439014.000
<input type="checkbox"/>	/ctolib/pmcgann/Invincible/sv_discovery_indexes/smrtp/CHS/ftp.sra.ebi.ac.uk/vol1/ERA562/ERA562I05/pacbio_hdf5/	m150916_081225_42196_c100828042550000001823180911251553_s1_p0.1.bax.h5	ctolib	jharkness	pmcgann	4731564083.000



# File and Object Data Aging Analytics

- Leverage time since access bucketing , visualization, and drill down search
- Default Bucketing

1 week	1 month	1 Quarter	1 year	1+ year
< 1 week	> 1 week ; < 1 month	> 1 month; < 3 months	> 3 months; < 1 year	>1 year

## Customize Bucket Ranges

### Modify Bucket

☒ 1 quarter

Between previous value and

3month

☒ 1 year

Between previous value and

1year

☒ 1 year+

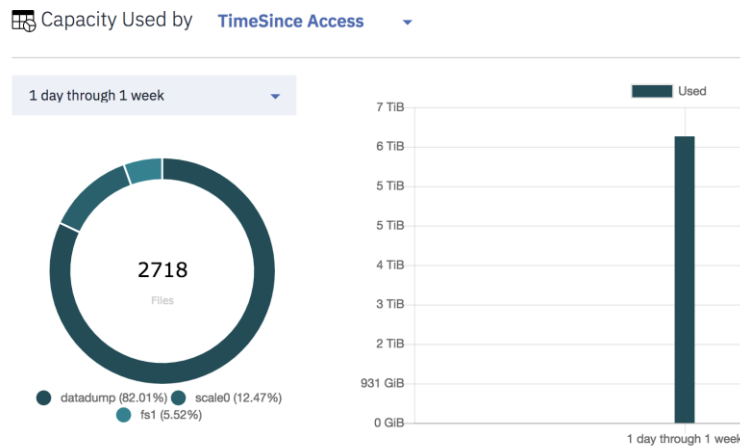
Greater than previous value

1year

Cancel

Submit

## Visualize Capacity Usage by size bucketing





# File and Object Size Distribution Analytics

## Time range drilldown search

←

timesinceaccess in ('4 year','2 -3 years')

Se

View results by: timesinceaccess

Results:

Generate Report

Add Tags

Convert to individual record mode.

▼

	timesinceaccess	Total Files	Total Size
<input type="checkbox"/>	4 year	937,530	0 Bytes
<input type="checkbox"/>	2 -3 years	9,025,937	166.9 TiB

Items per page: 20 | 1-2 of 2 items

1 of 1 pages

< 1 >

▼ TIMESINCEACCESS

☐ 2 -3 years ( 9,025,937 )

☐ 4 year ( 937,530 )

> OWNER

Results:

Generate Report

▼

Convert to individual records

Add Tags

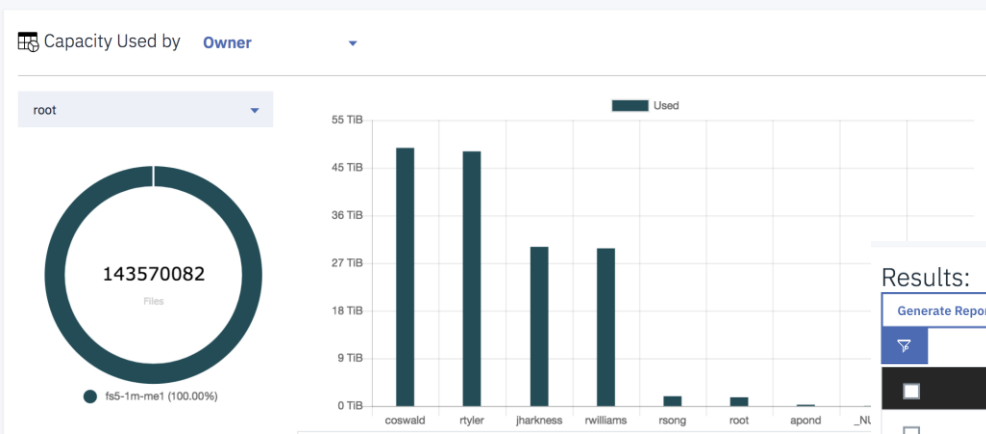
	path	filename	datasource	owner	fileset	atime	size
<input type="checkbox"/>	/ctolib/pmcgann/ThankYouAndGoodnight/rawPacbio/mnt/data3/vol56/2530572/0003/Analysis_Results/	runBlasr.2.out	ctolib	jharkness	pmcgann	2016-04-10T18:04:30.000Z	0.000
<input type="checkbox"/>	/ctolib/pmcgann/ThankYouAndGoodnight/rawPacbio/mnt/data3/vol56/2530574/0001/Analysis_Results/	runBlasr.0.out	ctolib	jharkness	pmcgann	2016-04-10T18:34:37.000Z	0.000
<input type="checkbox"/>	/ctolib/pmcgann/ThankYouAndGoodnight/rawPacbio/mnt/data3/vol56/2530574/0001/Analysis_Results/	runMakebam.0.out	ctolib	jharkness	pmcgann	2016-04-10T18:34:37.000Z	0.000
<input type="checkbox"/>	/ctolib/pmcgann/ThankYouAndGoodnight/rawPacbio/mnt/data3/vol56/2530574/0003/Analysis_Results/	runBlasr.2.out	ctolib	jharkness	pmcgann	2016-04-10T18:39:18.000Z	0.000





# File and Object Data Consumed by Owner

## Visualize Capacity Usage by Owner



## Visualize Capacity Usage by Owner

Results:

Generate Report

Add Tags

Convert to individual record mode.

	owner	Total Files	Total Size
<input type="checkbox"/>	rtyler	417	48.61 TiB
<input type="checkbox"/>	jharkness	8,981,376	30.39 TiB
<input type="checkbox"/>	apon	506	268.54 GiB
<input type="checkbox"/>	rwilliams	104,035	30.1 TiB
<input type="checkbox"/>	_NULL_	521	54.46 GiB
<input type="checkbox"/>	root	143,570,082	1.67 TiB
<input type="checkbox"/>	coswald	4,426	49.27 TiB
<input type="checkbox"/>	nobody	2	7.82 KiB
<input type="checkbox"/>	rsong	369,803	1.88 TiB

> PLATFORM

> SITE

> \_YEAR

> TEMPERATURE

> OWNER

☐ rsong ( 369,803 )

☐ root ( 143,570,082 )

☐ coswald ( 4,426 )

☐ nobody ( 2 )

☐ rtyler ( 417 )

☐ jharkness ( 8,981,376 )

☐ apond ( 506 )

☐ \_NULL\_ ( 521 )

☐ rwilliams ( 104,035 )




# Combine Criteria for Advanced Analytics

Example: space consumed by file size mapped against owner

View results by: size range owner

Results:

[Generate Report](#) [Add Tags](#) [Convert to individual record mode.](#)






<input type="checkbox"/>	size range	owner	Total Files	Total Size
<input type="checkbox"/>	small	root	72,130,399	1.1 TiB
<input type="checkbox"/>	extra small	jharkness	8,205,576	3.03 GiB
<input type="checkbox"/>	extra large	jharkness	4	7.32 TiB
<input type="checkbox"/>	medium	root	20	1.51 GiB
<input type="checkbox"/>	large	jharkness	3,607	20.56 TiB
<input type="checkbox"/>	extra small	rwilliams	40,107	31.46 MiB
<input type="checkbox"/>	large	root	15	582.58 GiB




# Combine Criteria for Advanced Analytics


Example: space consumed by file size and time since access mapped against owner

View results by: timesinceaccess sizorange owner

Results:



	timesinceaccess	sizorange	owner	Total Files	Total Size
<input type="checkbox"/>	1 year+	large	jharkness	3,607	20.56 TiB
<input type="checkbox"/>	1 year+	large	rsong	250	686.23 GiB
<input type="checkbox"/>	1 year+	extra small	jharkness	8,205,576	3.03 GiB
<input type="checkbox"/>	1 year+	extra small	rwilliams	40,107	31.46 MiB
<input type="checkbox"/>	1 year+	medium	_NULL_	310	54.46 GiB
<input type="checkbox"/>	1 year+	small	coswald	2	613.95 KiB



# Generate Custom Reports

Example: space consumed by file size and time since access mapped against owner

View results by: timesinceaccess sizorange owner

Results:

Generate Report Add Tags

**Generate Report**

Name

exampleReport

Current selected: 1  
Current report query: timesinceaccess IN ('1 year+') AND sizorange IN ('large') AND owner IN ('rtyler')

Group By: timesinceaccess sizorange owner

☒ View Individual Records

Cancel Submit

	timesinceaccess				Total Size
<input type="checkbox"/>	1 year+				20.56 TiB
<input type="checkbox"/>	1 year+				686.23 GiB
<input type="checkbox"/>	1 year+				3.03 GiB
<input type="checkbox"/>	1 year+				31.46 MiB
<input type="checkbox"/>	1 year+				54.46 GiB
<input type="checkbox"/>	1 year+	small	coswald	2	613.95 KiB
<input type="checkbox"/>	1 year+	large	root	15	582.58 GiB
<input checked="" type="checkbox"/>	1 year+	large	rtyler	417	48.61 TiB



# Map File and Object Data to Business Constructs

## Example: Tag data by project

### Policies

Add new policy.

inactive ☐ active

Name

project

### Schedule

☒ Now ☐ Daily ☐ Weekly ☐ Monthly

Policy Type

AUTOTAG

Filter

datasource='ctolib'

☒ Extract tag from path

Field

project

Depth

4

Example: root/folder1/subfolder2/subfolder3/subfolder4/...  
If depth is 4, Project = subfolder3



# Map File and Object Data to Business Constructs

Example: space consumed by file size and time since access mapped against owner and project

←

size in ('extra large','large','small','extra small','medium') AND timesinceaccess in ('1 quarter','1 year+') AND owner i... 

Search

View results by: 

sizein 

×

timesinceaccess 

×

owner 

×

project 

×

Results: 

Generate Report

Add Tags

Convert to individual record mode.

▼


	sizein	timesinceaccess	owner	project	Total Files	Total Size
<input type="checkbox"/>	extra small	1 year+	rsong	cgcc	15,174	29.74 MiB
<input type="checkbox"/>	large	1 year+	apond	fasta	16	44.72 GiB
<input type="checkbox"/>	extra small	1 year+	rwilliams	download	21,505	26.9 MiB
<input type="checkbox"/>	medium	1 year+	apond	star	25	377.83 MiB
<input type="checkbox"/>	medium	1 year+	jharkness	sv_discovery_indexes	628	4.16 GiB
<input type="checkbox"/>	extra small	1 year+	jharkness	fastq	2	188 Bytes
<input type="checkbox"/>	large	1 year+	apond	sv_discovery_indexes	1	2.38 GiB
<input type="checkbox"/>	large	1 year+	root		15	582.58 GiB



# Identify Potential Duplicate Data

Files with the same name and same size

## Dashboard Analytics

 Duplicate File Information

5,541,706

Duplicate Records

535.18 GiB

Total Capacity Consumed

Last Updated : 2019-04-07 00:14:41

## Search Results

Results:

[Generate Report](#)

[Add Tags](#)

[Convert to individual record mode.](#)



<input type="checkbox"/>	filename	size	Total Files	Total Size
<input type="checkbox"/>	.dummy	0	2	
<input type="checkbox"/>	.local-guid	14	9,543	130.47 KiB
<input type="checkbox"/>	000001.out	436	9	3.83 KiB
<input type="checkbox"/>	000001.out	439	15	6.43 KiB
<input type="checkbox"/>	000001.out	1243	5	6.07 KiB

## Command line duplicate data reports

- Duplicate data count
- Duplicate data capacity ordered by size



# File Type Distribution Report

## Command line file type distribution report

- Provides view of capacity consumed and count of files by file type grouped by datasource

```
usage: generate_report.py [-h] [-o filename] -u username [infile]
```

Generate Data Curation Reports

positional arguments:

infile                    Input file containing SQL query (default <stdin>)

optional arguments:

-h, --help                show this help message and exit

-o filename, --out filename

                          Output to file instead of stdout (will overwrite!)

-u username, --user username

                          User name with authority to create reports

```
python ./generate_report.py -u sdadmin sql/space_per_filetype.sql
```



# Leverage Default Capacity Showback Reports

Report Category	Report Description
Data aging reports	Provides insight about the age of files in the heterogeneous storage environment. Summary reports with count and capacity and detailed reports with full file details.
	Files accessed last 30 days, 31-60 days, 61-90 days, 91-180 days, 181-360 days, 361 -720 days, 720+ days
Size Snapshot	provides a view of the filesystem capacity, last access time, and last modify time
Space per collection	provides a view of the collection capacity, last access time, and last modify time
Space per user	provides a view of the capacity per user, last access time, and last modify time
Duplicates	Provides view of potential duplicate data – capacity and count for largest capacity data and capacity consumed and count of files by file type grouped by datasource
File type	Provides view of capacity consumed and count of files by file type grouped by datasource
Path Detail Report	Provides the amount of capacity consumed grouped by sub directory for the sub-directory depth specified by the user



# Data Governance, Content Inspection, and Content Classification with Spectrum Discover



# Medical center wanted to better manage research and clinical trial data

## Business challenge:

A large healthcare research center needed to address 4 key elements: 1) catalog large genomic reference dataset 2) monitor and report on data location 3) finding PHI/PII data from genomic and medical imaging datasets 4) establishing data usage patterns

## Outcome:

Customer is rolling out 30PB of Spectrum Discover that is being used to analyze and develop more use cases and insight into the 100+ PB of medical data currently stored online.

30PB

of initial data is managed by Spectrum Discover

Identify personal data

find PHI/PII from medical imaging datasets

Identify data usage

establish patterns for monitoring and reporting





# Metadata tags

Define and apply custom tags according to customer defined data governance taxonomy to manage unstructured data on premises and in the cloud



Create custom metadata field names and tags

- Unique to organizational schema/taxonomy
- Manual and/or via API for automated insertion

Enables organizations to describe data with more meaningful tags

Metadata tags can be Open or Restricted

- Open tags allow user to specify value of their choice
- Restricted tags enforce only defined values to be used

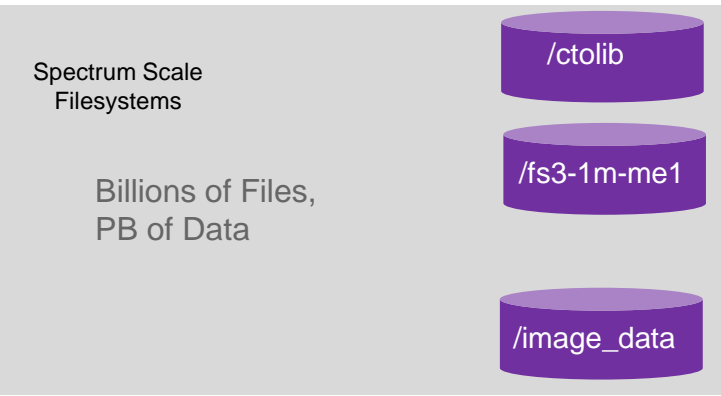
Field Name	Type	Tags	Edit/Delete
COLLECTION	Open		
TEMPERATURE	Open		
project	Open		
department	Open		
classification	Restricted	<span>public</span> <span>confidential</span> <span>sensitive</span>	

20 items per page | 1 of 5 items



# Use Case: Curating the Research Data for Placement Optimization (Data Governance)

User	Department Tag	Project Tag	Project State Tag	Spectrum Scale Fileset / Base Directory
ibmuser1	staff	phase1	active	/whole_cell
ibmuser2	postdoctoral	phase2	inactive	/nucleus
ibmuser3		phase3	active	/polysomes



1. Collect technical metadata (file name, size, etc)



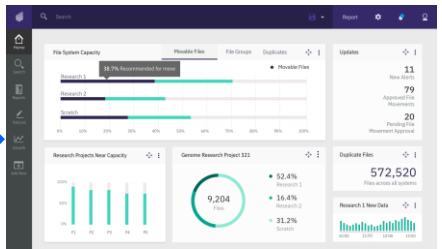
2. Policy-based auto-tag (enriching data with customer specific tags)

Capacity Reporting

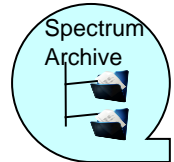


4. Generate reports (Capacity show back)

3. Ad-hoc filtered search



5. Move to tape





# 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

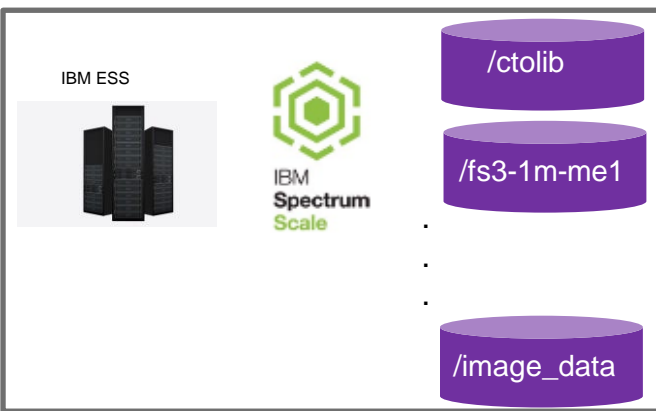
Automates the identification and classification of documents that could potentially contain Personally Identifiable Information (PII) and Sensitive Data.

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

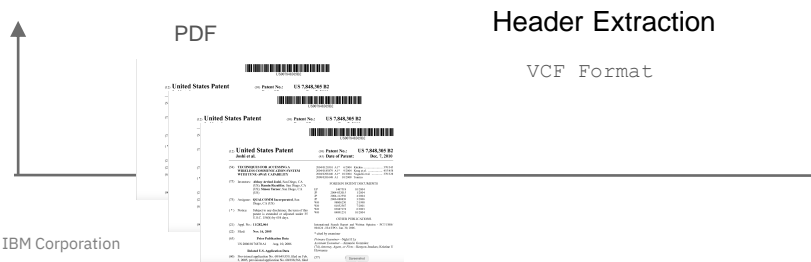
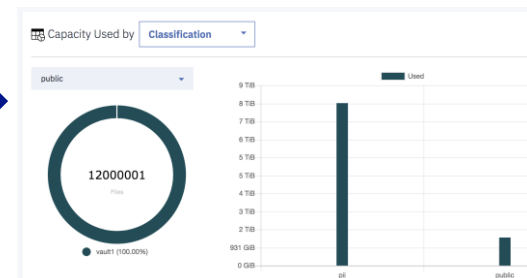


# Demo Use Case: Patent Inspection

Show me where my data  
resides for patents filed in  
2016



System  
Metadata  
Events





# Demo Use Case: Patent Inspection



US007848305B2

(12) **United States Patent**  
**Joshi et al.**

(10) **Patent No.:** **US 7,848,305 B2**  
(45) **Date of Patent:** **Dec. 7, 2010**

(54) **TECHNIQUES FOR ACCESSING A  
WIRELESS COMMUNICATION SYSTEM  
WITH TUNE-AWAY CAPABILITY**

(75) Inventors: **Abhay Arvind Joshi**, San Diego, CA  
(US); **Ramin Rezailifar**, San Diego, CA  
(US); **Simon Turner**, San Diego, CA  
(US)

(73) Assignee: **QUALCOMM Incorporated**, San  
Diego, CA (US)

(\*) Notice: Subject to any disclaimer, the term of this  
patent is extended or adjusted under 35  
U.S.C. 154(b) by 654 days.

(21) Appl. No.: **11/282,064**

(22) Filed: **Nov. 16, 2005**

(65) **Prior Publication Data**  
US 2006/0176870 A1 Aug. 10, 2006

## **Related U.S. Application Data**

(60) Provisional application No. 60/649,959, filed on Feb.  
3, 2005, provisional application No. 60/698,566, filed  
on Jul. 12, 2005

2004/0120301 A1 \* 6/2004 Kitchin ..... 370/345  
2004/0185879 A1 \* 9/2004 Kong et al. .... 455/458  
2004/0208140 A1 \* 10/2004 Noguchi et al. .... 370/328  
2008/0261648 A1 10/2008 Tomizu

## **FOREIGN PATENT DOCUMENTS**

EP	1467518	10/2004
JP	2004-032015	1/2004
JP	2004-112556	4/2004
JP	2006-080839	3/2006
WO	98006230	2/1998
WO	01052567	7/2001
WO	03047174	6/2003
WO	04091231	10/2004

## **OTHER PUBLICATIONS**

International Search Report and Written Opinion - PCT/US06/  
004124 - ISA/EPO - Jun. 30, 2006.

\* cited by examiner

*Primary Examiner*—Nghi H Ly

*Assistant Examiner*—Amancio Gonzalez

(74) *Attorney, Agent, or Firm*—Kenyon Jenckes; Kristine U  
Ekwueme

(57)

Screenshot





IBM  
Spectrum  
Discover

# IBM Spectrum Discover – Content Classification Workflow

Policy based, fully automated content inspection and classification

- 1. Leverage pre-configured terms / regular expressions and classification mappings
- 2. Create custom terms / regular expressions
- 3. Modify classification mappings

PII	Sensitive	Public
SSN	Confidential	Security: None
DOB		
Phone #		



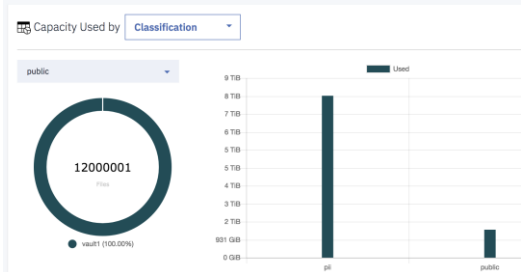
What kind of sensitive data is stored in my enterprise?

Need to identify all the documents containing Credit Card Info, SSN, Emails and Phone Numbers.

Where is this data stored?



System metadata

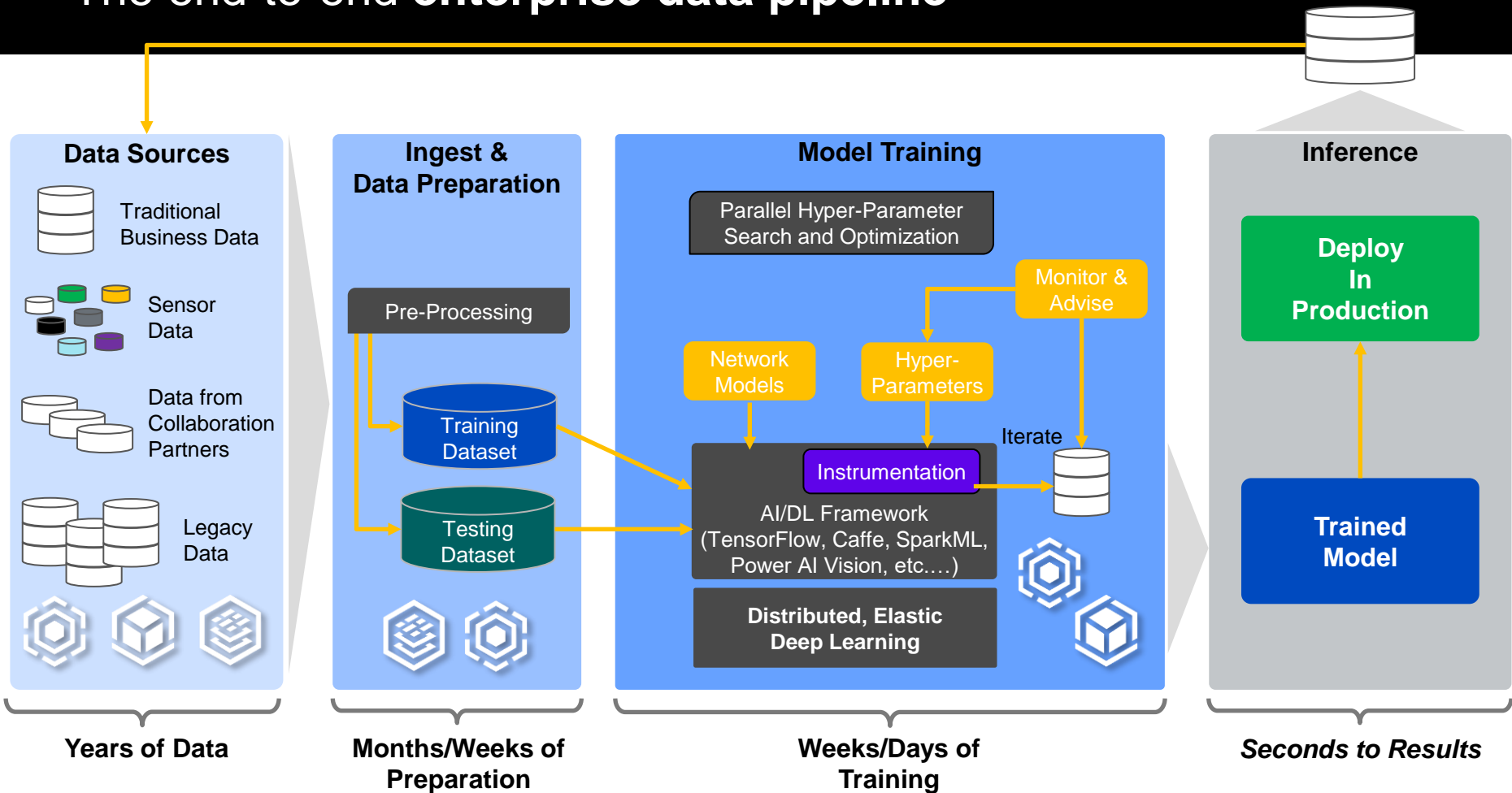




# Spectrum Discover and the AI Pipeline



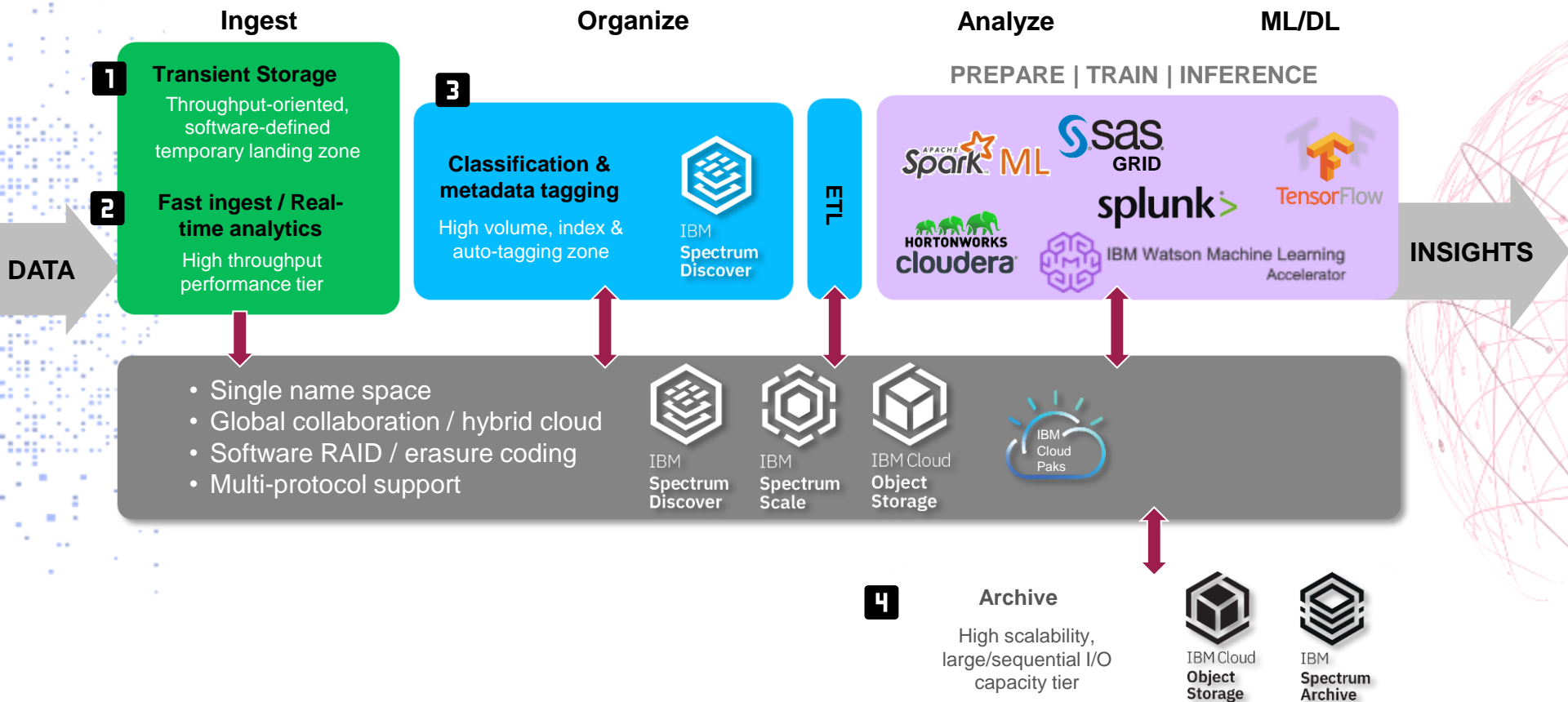
# The end-to-end enterprise data pipeline





# IBM Spectrum Storage for the AI data pipeline

*The fastest path from ingest to insights*





# Spectrum Discover + Power AI Vision Use Cases and Customer Example

## Primary Use Cases

1. Accelerating data curation and acquisition of training data sets for Power AI Vision from heterogeneous data sources on premises and in the cloud
2. Event driven AI pipeline to automatically classify and catalog newly ingested IOT data using Power AI Vision inference models

#powerai-vision-team 

☆ | 👤 554 | 📌 2 | Discussion/Troubleshooting - IBM PowerAI Vision.



Anybody got any experience of handling vast volumes of image data and how it interacts with AI Vision? In my scenario, I will have **20 sites, 6 production lines** and **50 models in each production line**. I will **store all this data centrally**, it **will be enriched with metadata including xml overlays for boundary boxes and hierarchy (site, production line, model....)** Might be as simple as a filesystem with the hierarchy (site, production line, model) in, but how to make that simple to configure, flexible and help structure my data so I can automate sending it to AI Vision as a folder ready for training.

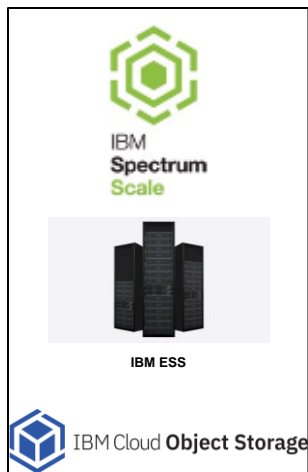
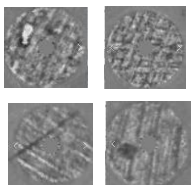


# Use Case: Automated Wafer Manufacturing Image Classification

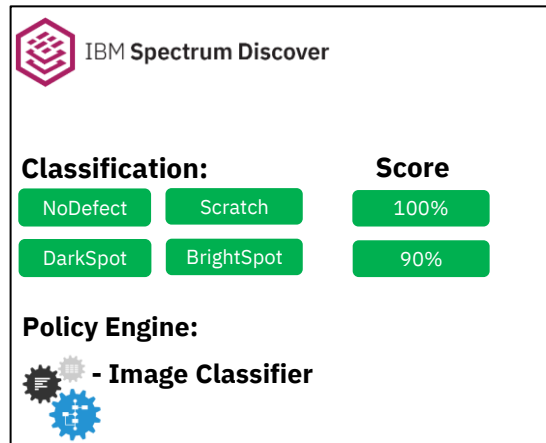
Event driven architecture to automatically classify and catalog wafer manufacturing data using PowerAI Vision inference model, Spectrum Discover, and Spectrum Scale / ESS / COS

1. New imaging data ingested into Spectrum Scale / ESS, IBM COS storage
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 PowerAI Vision classification model, captures results and indexes into Spectrum Discover

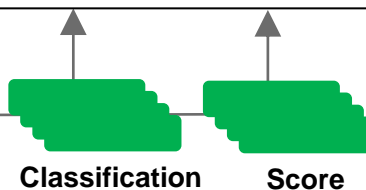
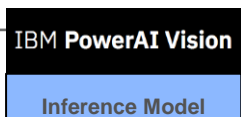
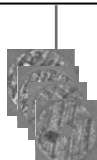
## 1. Wafer Manufacturing Imaging Data Ingestion



## 2. System Metadata Events

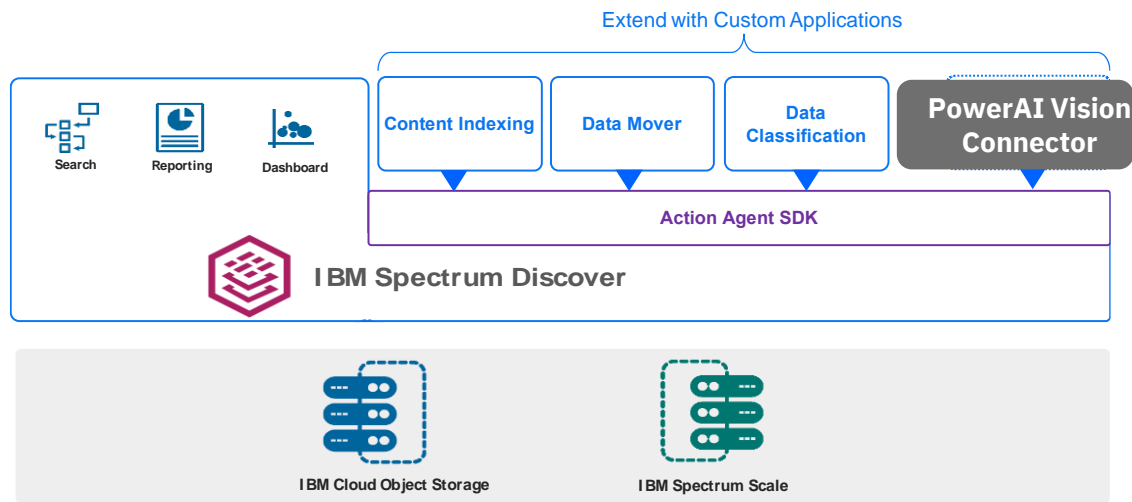


## 3. Image classification workflow





# Spectrum Discover – PowerAI Vision Application Plugin



## Power AI Vision Connector

### Spectrum Discover Application

Reads images from Spectrum Scale and / or COS, passes to Power AI Vision inference model, captures classification and score output, and updates Spectrum Discover catalog with results

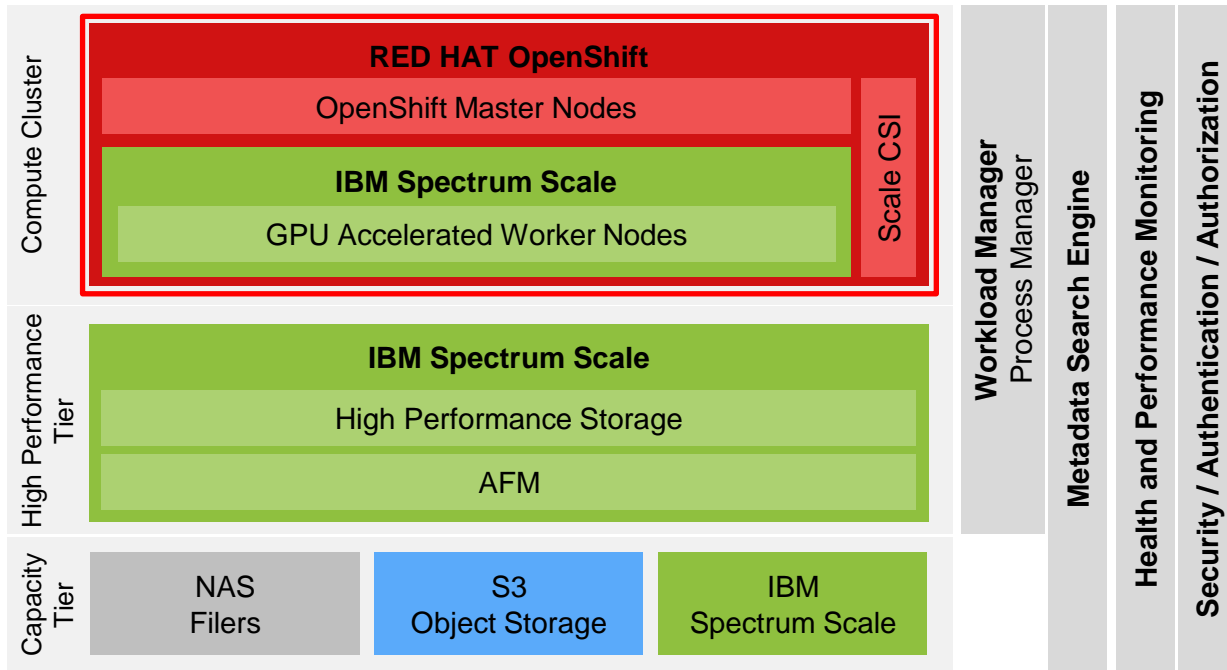


 [IBM / Spectrum\\_Discover\\_App\\_Catalog](#)



# Data Accelerator for AI and Analytics (DAAA)

How can I easily / efficiently provision, test, deploy and scale my containerized workloads?





# Data Accelerator for AI and Analytics RedPaper

## Data Accelerator for AI and Analytics

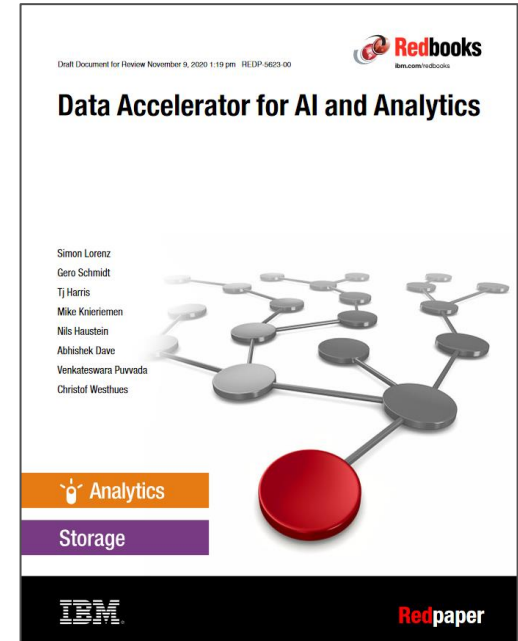
Watch:

<https://www.spectrumscaleug.org/event/ssugdigital-data-accelerator-for-analytics-and-ai-daaa/>

Read:

<http://www.redbooks.ibm.com/redpieces/abstracts/redp5623.html>

(published November 09, 2020)





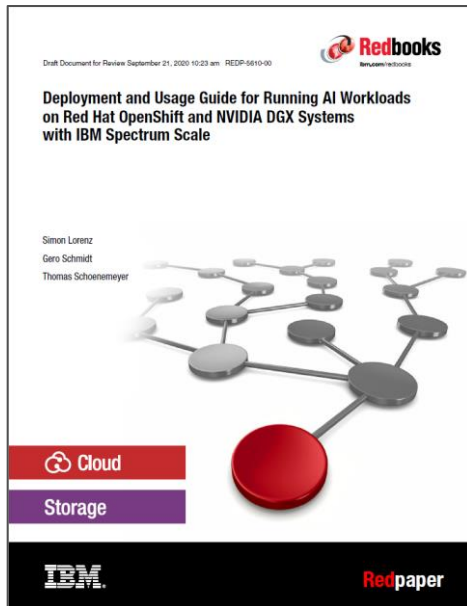
# AI Example Use Case: Autonomous Driving



Worked on an IBM Redpaper:

## Deployment and Usage Guide for Running AI Workloads on Red Hat OpenShift and NVIDIA DGX Systems with IBM Spectrum Scale

Visit:

<http://www.redbooks.ibm.com/redpieces/abstracts/redp5610.html>



 <b>Spectrum Scale Expert Talks</b>	
<b>Episode 8:</b> <b>Multi-node scaling of AI workloads using NVIDIA DGX, OpenShift and Spectrum Scale</b>	
<b>Show notes:</b> <a href="http://www.spectrumscaleug.org/experttalks">www.spectrumscaleug.org/experttalks</a>	<b>Join our conversation:</b> <a href="http://www.spectrumscaleug.org/join">www.spectrumscaleug.org/join</a>





# Extend the functionality of Spectrum Discover

*with Spectrum Discover Application Catalog*

## Community-supported catalog of open source Action Agents

- Enhance the capabilities of Spectrum Discover with third-party extensions
- Find and install available extensions via CLI (with Docker Hub)
- Develop and share new extensions, supported with sample code and a fully-published API



IBM  
**Spectrum**  
**Discover**



# Nvidia DGX2 and Figure Eight Wildfire Dataset with Spectrum Discover



# How is ai helping fire fighting?

- Fire fighting leaders on the ground and consultants miles away can now improve fire fighting safety using **UAS based AI cameras** and other navigational tools to see in real-time.
- **Airborne lidar**, lets researchers visualize trees in 3D, supplemented with ground-based lidar, which details the vegetation underneath the trees.
- **AUDREY, an AI Fire Fighting Assistant**

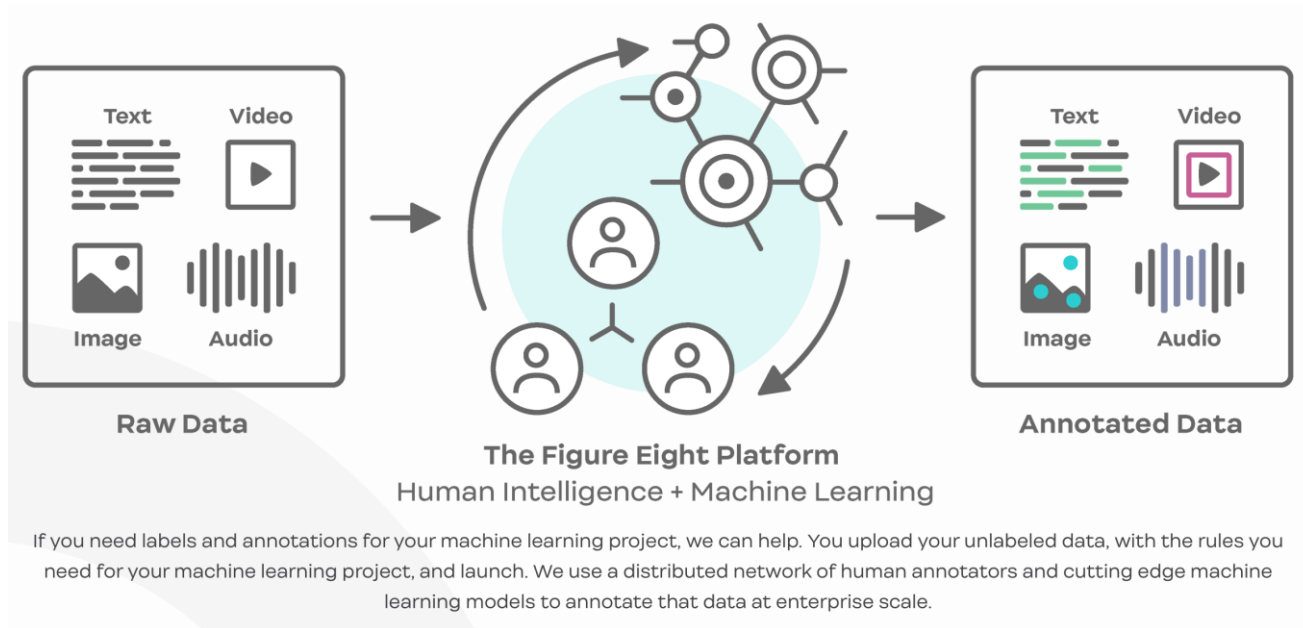
being taught fire behavior and the risks firefighters face to assist firefighters, incident managers, and dispatchers to keep personnel safe.





## Dataset Annotation with Figure-eight

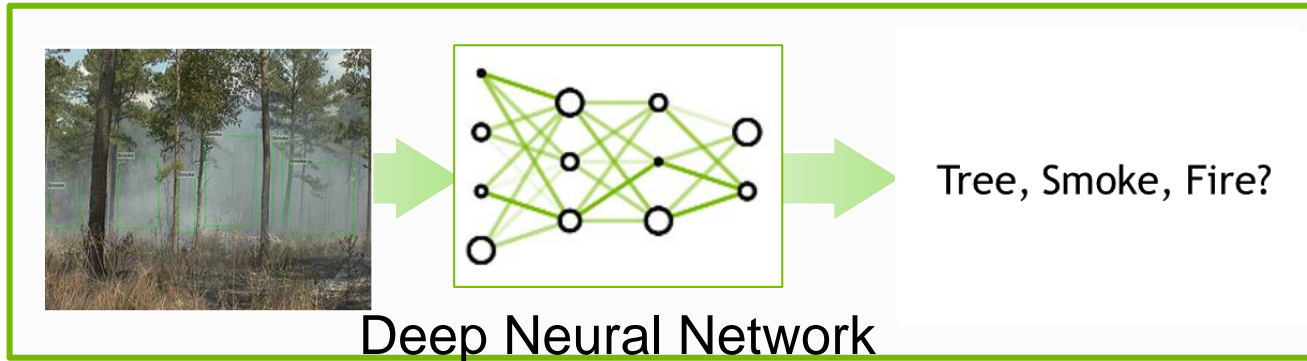
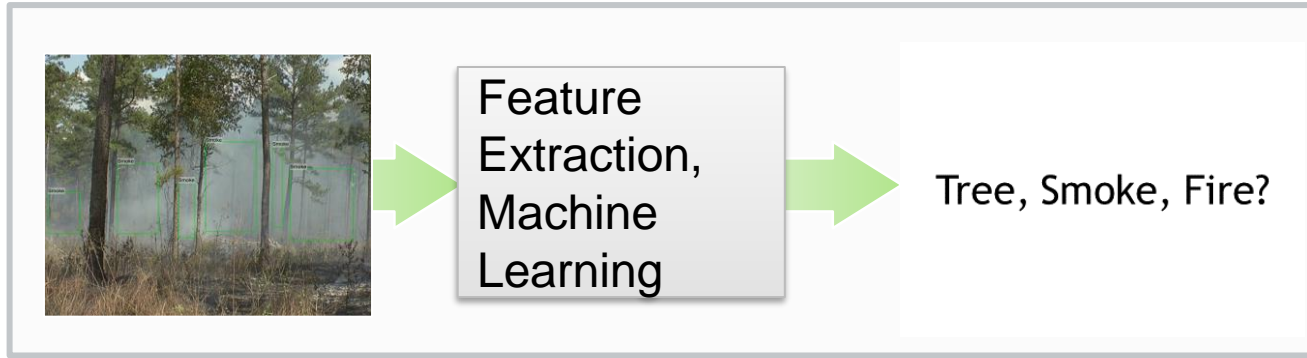
figure eight





# IBM Wildfire Dataset

Identify smoke at fire line for prescribed burns

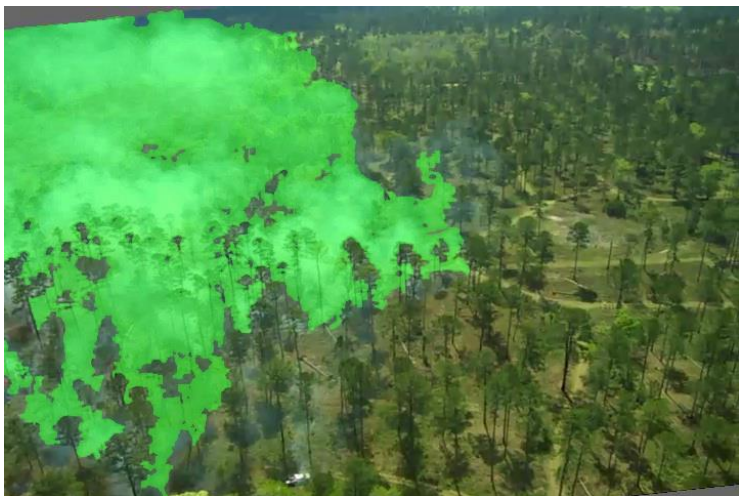




# How did Figure-eight annotate the wildfire dataset

## Semantic Segmentation

Pixel-level labeling for computer vision projects



Ontology

display_color	description	class_name	output_value
#21ff17		Smoke	1

filename	type	id	category	maskfile	visibility
video_261_0047987.jpg	mask	261_0047987	Smoke	mask_261_0047987.jpg	visible
video_261_0047988.jpg	mask	261_0047988	Smoke	mask_261_0047988.jpg	visible
video_261_0047989.jpg	mask	261_0047989	Smoke	mask_261_0047989.jpg	visible

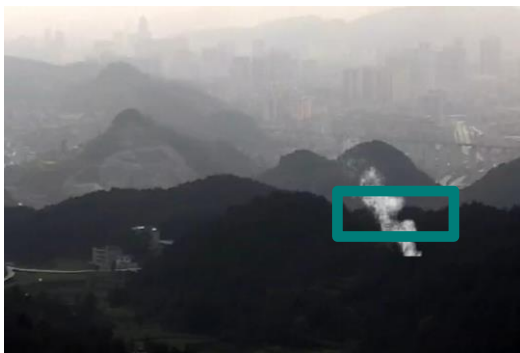
_id	_started_at	_tainted	_channel	_trust	_worker_id	_country	_region	_city	_ip	_annotation	_image_broke	_image_broke	_image_url					
4925946530	6/27/19 21:57	FALSE	cf_internal	1	45179252	USA	CA	San Bruno	12.248.233.5	{"url":"https://nickfigure8.s3.amazonaws.com/Nvidia%20pics/video_261_0047987.jpg"	FALSE	FALSE	http://nickfigure8.s3.amazonaws.com/Nvidia%20pics/video_261_0047987.jpg					
4926306881	6/28/19 1:30	FALSE	cf_internal	1	45179252	USA	CA	Livermore	76.103.23.22	{"url":"https://nickfigure8.s3.amazonaws.com/Nvidia%20pics/video_261_0047988.jpg"	FALSE	FALSE	http://nickfigure8.s3.amazonaws.com/Nvidia%20pics/video_261_0047988.jpg					
4925946539	6/27/19 21:57	FALSE	cf_internal	1	45179252	USA	CA	San Bruno	12.248.233.5	{"url":"https://nickfigure8.s3.amazonaws.com/Nvidia%20pics/video_261_0047989.jpg"	FALSE	FALSE	http://nickfigure8.s3.amazonaws.com/Nvidia%20pics/video_261_0047989.jpg					



## How did Figure-eight annotate the wildfire dataset

### Bounding Box Object Detection

Polygon based bounding box annotations on wildfire dataset

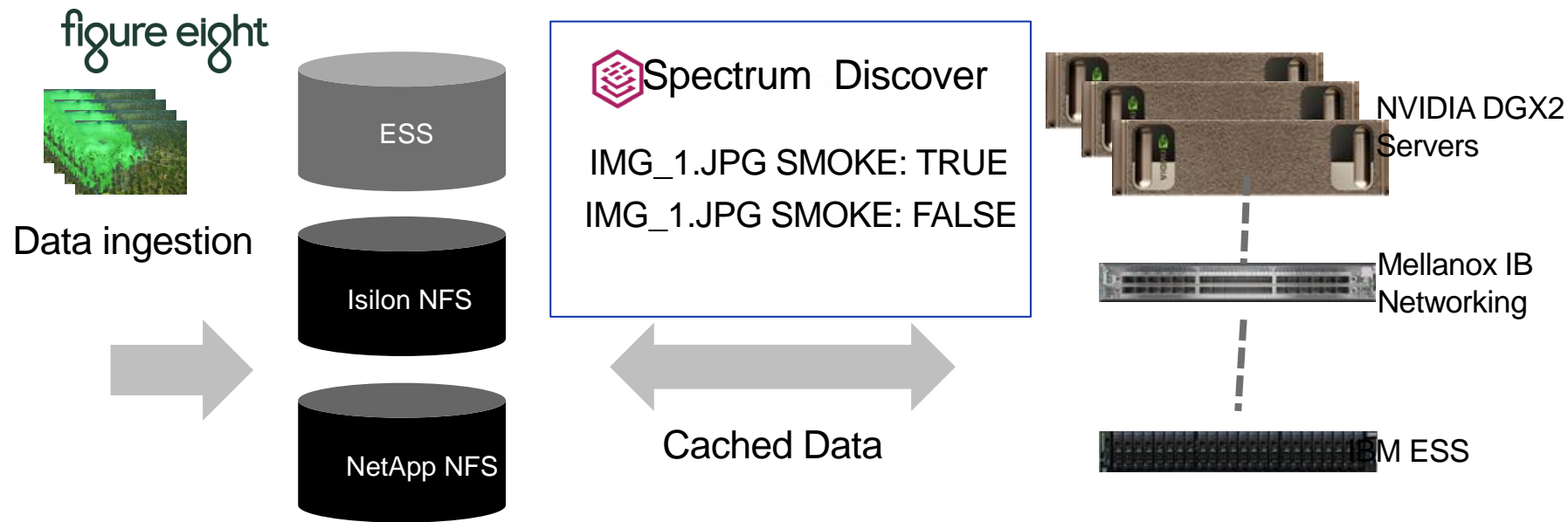


filename	type	id	category	annotated_b_x	y	height	width	visibility
0001.png	box	2bf4e928-67	Smoke	human	418	397	147	50 visible
0001.png	box	3ef01461-10	Smoke	human	628	447	99	74 visible
0001.png	box	36ad3edb-34	Smoke	human	784	380	176	95 visible
0001.png	box	4d212cf3-0a	Smoke	machine	559	483	34	48 hidden
0001.png	box	0c932ce3-f8	Smoke	machine	1214	464	113	23 hidden
0001.png	box	2cbf8d43-35	Smoke	machine	1773	313	57	57 hidden
0002.png	box	2bf4e928-67	Smoke	machine	418	397	147	50 visible
0002.png	box	3ef01461-10	Smoke	machine	628	447	99	74 visible
0002.png	box	36ad3edb-34	Smoke	machine	784	380	176	95 visible
0002.png	box	743464c4-e5	Smoke	machine	992	355	195	107 visible



## *File caching/prefetching w/Spectrum Discover leveraging Figure-eight annotations*

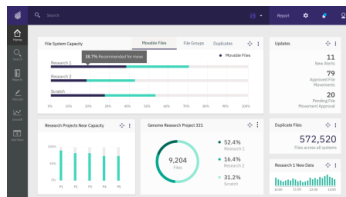
1. Annotated dataset by Figure-eight for IBM Fire project loaded into warm tier(s)
2. File metadata and annotations performed by Figure-eight indexed into Sp. Discover catalog
3. Data scientist leverages Sp. Discover to search for data leveraging index of Figure-eight annotations and triggers caching the matching data to an ESS / Spectrum Scale high performance tier
4. Run TensorFlow job and capture new annotations metadata into Sp. Discover





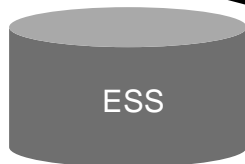
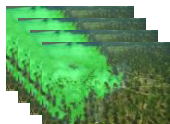
# Filtered Caching with Spectrum Discover Based on Labels / Annotations

Cache images from San Bruno that have smoke in them

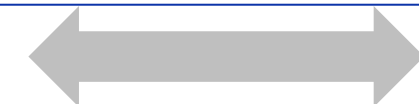


Datasource	Path	Filename	Country	Region	City	Category	Visibility
ESS-gpfs0	/gpfs0/ibm_fire/Nvidia%20pics/	video_261_0047987.jpg	USA	CA	San Bruno	Smoke	visible
ESS-gpfs0	/gpfs0/ibm_fire/Nvidia%20pics/	video_261_0047988.jpg	USA	CA	Livermore	Smoke	visible
ESS-gpfs0	gpfs0/ibm_fire/Nvidia%20pics/	video_261_0047989.jpg	USA	CA	San Bruno	Smoke	not visible

figure eight



 Spectrum Discover



Cached Data:  
video\_261\_0047987.jpg



NVIDIA DGX2  
Servers



Mellanox IB  
Networking



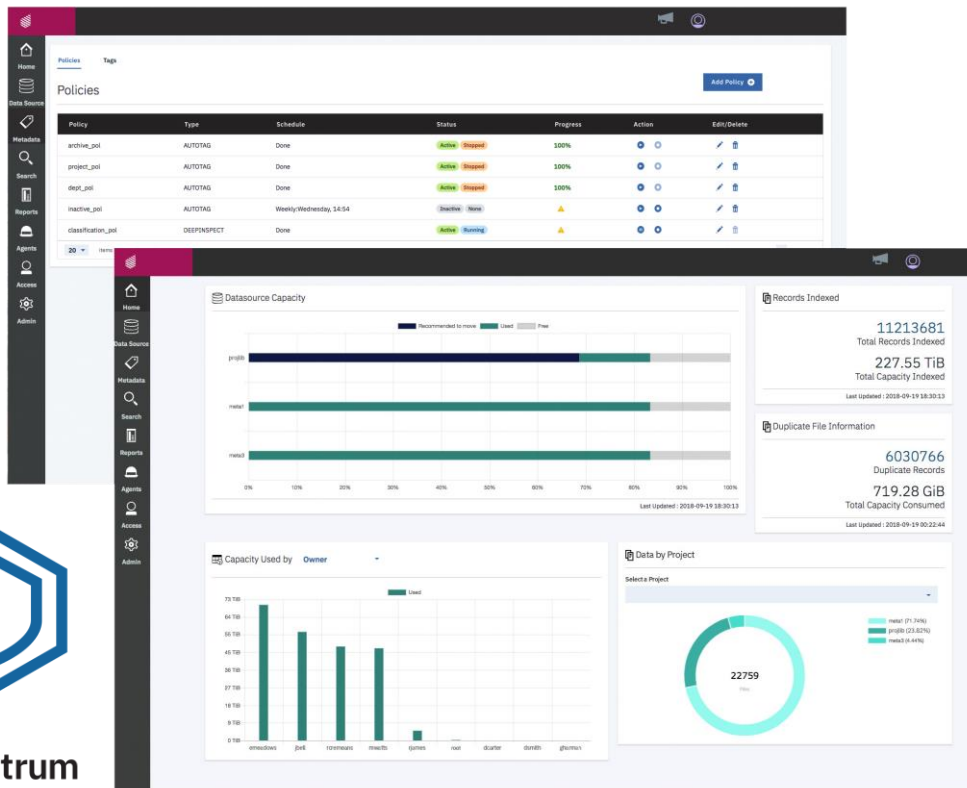
IBM ESS



Spectrum Discover Support for Spectrum Protect



# Gain deep insights into data in backup environments *with support for IBM Spectrum Protect*



- Gain Deep Insights into Data in Backup Environments
- Easily connect to Spectrum Protect to discover, index, and label files in backups
- Quickly find and activate cold data in backup/archive for analytics and AI
- Cleanup Spectrum Protect environment for better storage utilization



IBM  
Spectrum  
Protect



# 5 Common data protection questions

1 Do I have data in backup pools that have aged and could be moved to archive?

2 Do I have abandoned data and / or dark data in my Spectrum Protect environment?

3 What types of data am I backing up and how big is it? Is there data that I can remove from backup?

4 What is the content of my active and inactive data?

5 How can I map this information against organizational constructs / custom tags?



# Integration – Get started today

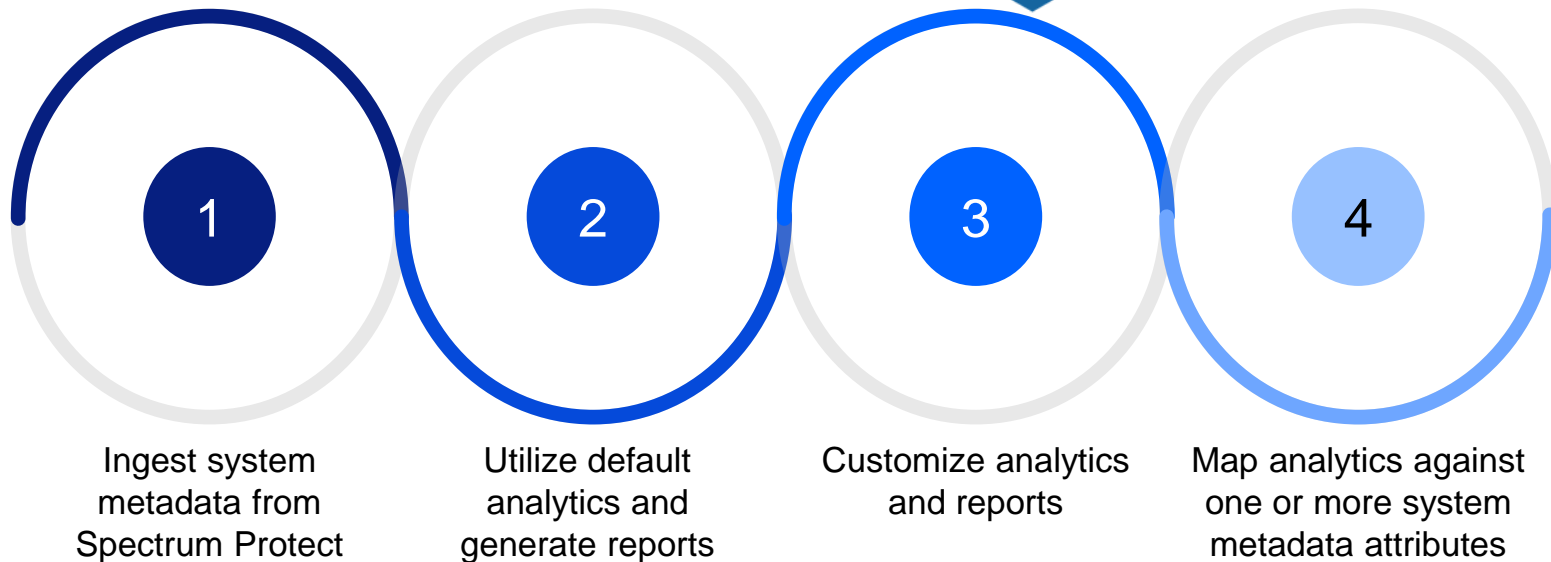
Spectrum Discover with built-in analytics



IBM **Spectrum Discover**



IBM **Spectrum Protect**



Find



Curate and organize



Inspect and classify



Optimize





# “Bucketize your data

The image shows two overlapping 'Modify Bucket' configuration windows. The background window is for 'Age' configuration, with a red circle around the '1 quarter' option and a line pointing to the 'Age' label. The foreground window is for 'Size' configuration, with a red circle around the 'small' option and a line pointing to the 'Size' label. Both windows have a 'Submit' button.

**Age Configuration:**

- ☒ 1 quarter
- Between previous value and: 3 month
- ☒ 1 year
- Between previous value and:
- ☒ 1 year+
- Greater than previous value

**Size Configuration:**

Please make sure that the maximum value for each bucket is greater than the value assigned to the previous bucket

Bucket Name

SizeRange

- ☒ extrasmall
- Values less than: 4 KIB
- ☒ small
- Between previous value: 1 MiB

Cancel Submit

## Age analytics

- How long has the data been sitting around?
- What data can I move from a backup pool to archive or delete?

## Type analytics

- What type of data do I have?
- Where is ? data type located?

## Size analytics

- How big are the files in my backup set?
- How big is my backup set?

## Filespace analytics

- How much data is being stored in each file space?



# Example – Search Visualization (cont'd)

## Data age, mapped to other characteristics

Select specific criteria for further analysis

The screenshot shows a search visualization interface with several filter panels. Red circles highlight the following criteria:

- DATASOURCE:** ☒ GTK-SP-0000 (154,739,240)
- PLATFORM:** ☒ IBM Spectrum Protect (159,783,036)
- NODENAME:** ☒ DESKTOP-LSM732T (25), ☒ GTK-META-0008.TUC.STGLABS.IBM.COM (35,759,213), ☒ GTK-SP-0000.TUC.STGLABS.IBM.COM (118,265,163), ☒ HAN.TUC.STGLABS.IBM.COM (101,973), ☒ LEIA.TUC.STGLABS.IBM.COM (612,945), ☒ LIVE.TUC.STGLABS.IBM.COM (1,896,000)
- TIMESINCEACCESS:** ☒ 1 year (159,876,771)
- FILESIZE:** ☒ / (117), ☒ /gifs/gpfs0 (716,804), ☒ /home (229), ☒ /mnt/beast (118,265,150), ☒ /mnt/isilon (35,759,172), ☒ \\9.11.201.171\ifs\discover\data\gtk (2,123)
- SIZE RANGE:** ☒ large (1,638)

A red circle with a right arrow and the text "Next arrow" is located to the right of the filter panels.

**Search Visualization Summary:**

Criteria	Count
GTK-SP-0000 (DATASOURCE)	154,739,240
IBM Spectrum Protect (PLATFORM)	159,783,036
DESKTOP-LSM732T (NODENAME)	25
GTK-META-0008.TUC.STGLABS.IBM.COM (NODENAME)	35,759,213
GTK-SP-0000.TUC.STGLABS.IBM.COM (NODENAME)	118,265,163
HAN.TUC.STGLABS.IBM.COM (NODENAME)	101,973
LEIA.TUC.STGLABS.IBM.COM (NODENAME)	612,945
LIVE.TUC.STGLABS.IBM.COM (NODENAME)	1,896,000
1 year (TIMESINCEACCESS)	159,876,771
/ (FILESIZE)	117
/gifs/gpfs0 (FILESIZE)	716,804
/home (FILESIZE)	229
/mnt/beast (FILESIZE)	118,265,150
/mnt/isilon (FILESIZE)	35,759,172
\\9.11.201.171\ifs\discover\data\gtk (FILESIZE)	2,123
large (SIZE RANGE)	1,638

**Next arrow**

**Show me large files over 1 year old for a particular Spectrum Protect server and all associated nodenames**



# Example – Ad hoc search

or start a visual exploration

☐ Cluster

☐ Platform

☐ SizeRange

☐ NodeName

☐ State

☐ Project

☐ Datasource

☐ Site

☐ TimeSinceAccess

☐ Fileset

☐ COLLECTION


☒ Owner

☐ Tier

☐ MgmtClass

☐ Filespace

☐ TEMPERATURE

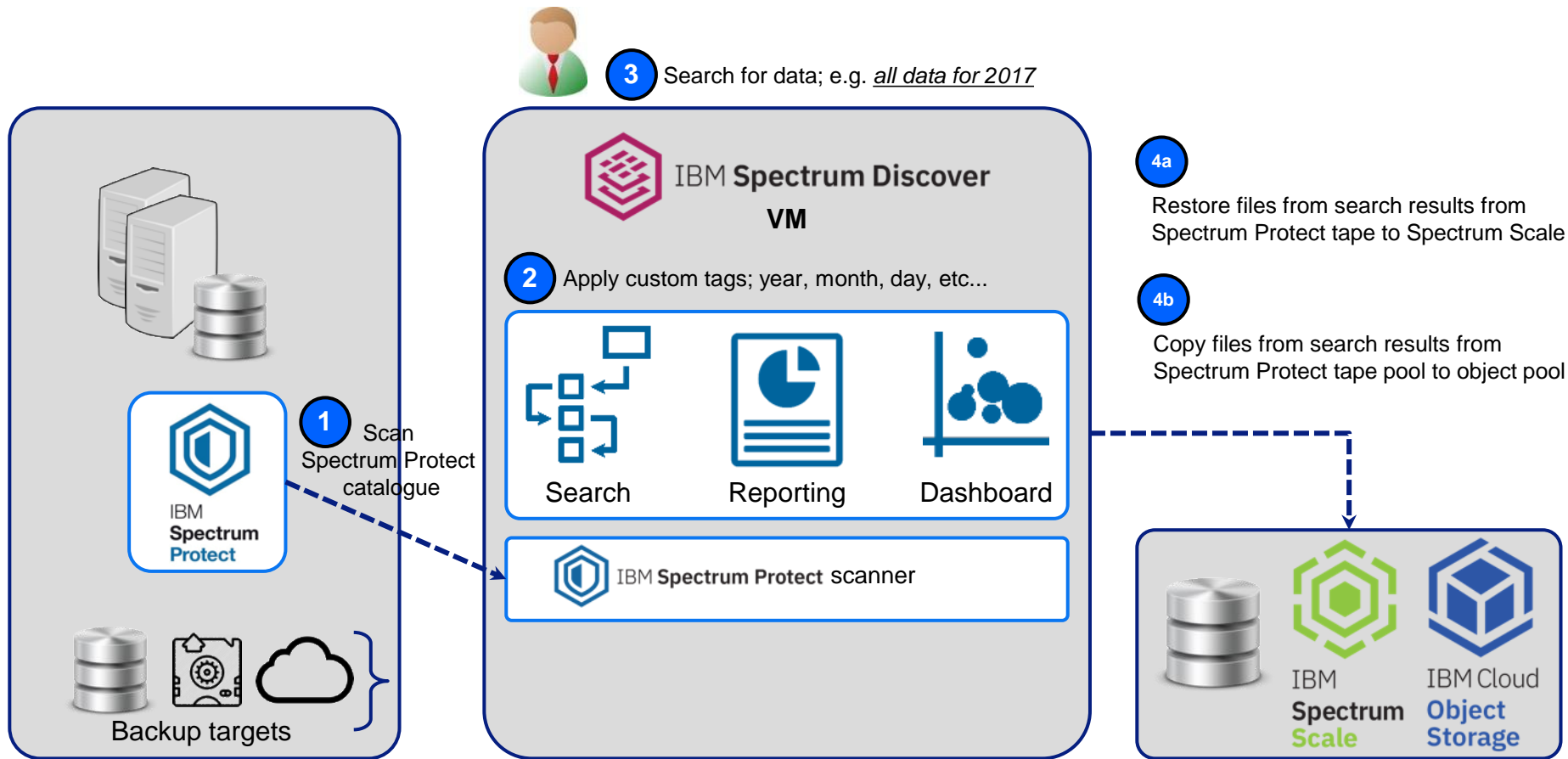


## Search Examples:

- Show me all data owned by user abc
- Show me all data from nodename xyz
- Show me all data from nodename xyz backed up in 2017
- Show me all data for a particular project



# Scan, tag, search, move...







# Lineage and Provenance





# Data Provenance and Lineage for Analytics

*Scientific Research is generally held to be of good provenance when it is documented in detail sufficient to allow **reproducibility**.*

*Deductions and Inferences are **reliable** when the processes used to create them are **reproducible**.*

- “If this data could talk”, Margo Seltzer et al., 2017
- “Ensuring reliable datasets for environmental models and forecasts. Ecological Informatics”, Boose et al., 2007

Spectrum Discover  
Provenance and Data Lineage  
will assist scientists to  
**track** their data through all **transformations**,  
**analyses**, and **interpretations**.

Make analytical models accountable!





# Facets of Data Lineage and Provenance

## Origin

- Where did this data come from?
- What dataset was used to derive this result?
- What sampling frequency?
- What drugs were administered at the time of this trial?

## Transformations

- What algorithms were used?
- What transformations were applied?
- What sampling was used?
- How many iterations?
- What cleanup filters were used?
- How many data points were discarded?

## Reproducibility

- If I have the same input and I run the same model, would I derive the same conclusion or inference?
- Can I trust the result published in this paper?
- **How do I know my analytics were not tampered with?**

## Challenges

### Manual cataloging is

Inconsistent  
Incomplete  
Lacks formalism  
Cross group collaboration is prohibitively cumbersome

### Team member churn

### Hours of wasted

### analysis, compute

### Dark Data – Wasted storage space

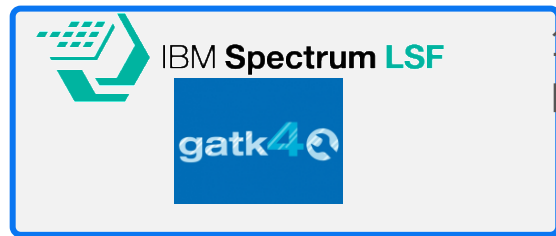
### Documented Evidence incomplete

### Lack of Trust !!!

The Goal:  
Make  
analytics  
accountable!

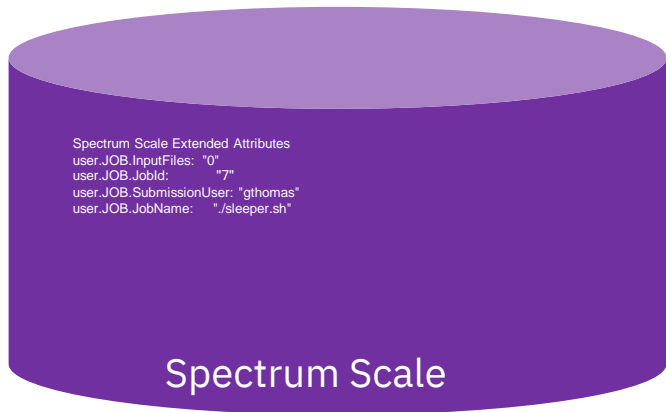


# Lineage and Provenance AI Solution Blueprint



1. Run GATK4 pipeline

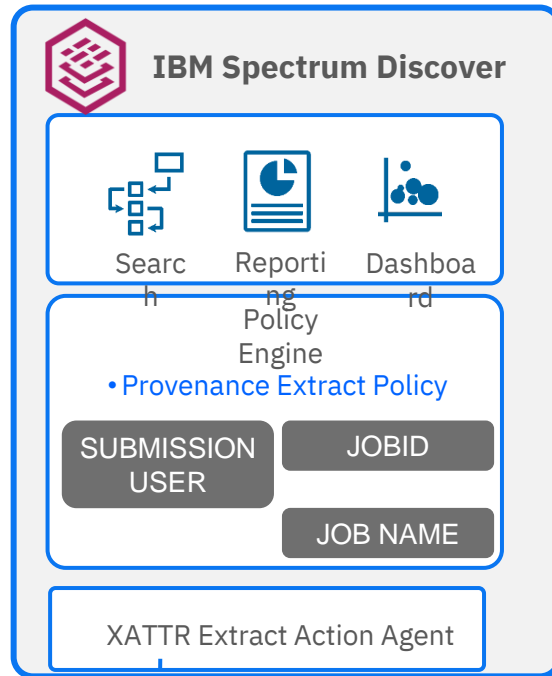
2. Write Provenance data as XATTRS



3. Ingest system metadata

5. Leverage Provenance info

Data Scientist



4. Extract and index provenance info



# Spectrum LSF Provenance Data

The extended attributes are the ones starting with "user.JOB"

```
[gthomas@p95a07 ~]$ pwd
/home/gthomas
[gthomas@p95a07 ~]$ bsub -q gatk -o /gpfs/gpfs_2mb/gilbert/out2.txt -Ep
~/lsf/10.1/misc/examples/data_prov/tag.sh ./sleeper.sh
[gthomas@p95a07 ~]$ mmlsattr -d -L /gpfs/gpfs_2mb/gilbert/out2.txt
file name:                /gpfs/gpfs_2mb/gilbert/out2.txt
metadata replication: 1 max 2
data replication:        1 max 2
immutable:               no
appendOnly:              no
flags:
storage pool name:      data
fileset name:           root
snapshot name:
creation time:          Mon May  6 22:57:56 2019
Misc attributes:        ARCHIVE
Encrypted:               no
user.JOB.InputFiles:    "0"
user.JOB.JobId:         "7"
user.JOB.SubmissionUser: "gthomas"
user.JOB.JobName:       "./sleeper.sh"
user.JOB.Status:        "64"
user.JOB.StartTime:     "1557198405"
user.JOB.FinishTime:    "1557198496"
user.JOB.SubmissionCmd: "./sleeper.sh"
user.JOB.JobWorkDir:    "/home/gthomas/"
```



# Spectrum LSF Provenance Data with Spectrum Discover

## Search Based on custom tags for data provenance

Discover what's in your Data

Search

or start a visual exploration

- |  |  |   |
|--|--|---|
| <input type="checkbox"/> Cluster               | <input type="checkbox"/> Datasource                | <input type="checkbox"/> Owner                            |
| <input type="checkbox"/> Platform              | <input type="checkbox"/> Site                      | <input type="checkbox"/> Tier                             |
| <input type="checkbox"/> SizeRange             | <input type="checkbox"/> TimeSinceAccess           | <input type="checkbox"/> COLLECTION                       |
| <input type="checkbox"/> TEMPERATURE           | <input type="checkbox"/> Test3                     | <input checked="" type="checkbox"/> JOB_InputFiles        |
| <input checked="" type="checkbox"/> JOB_JobId  | <input checked="" type="checkbox"/> JOB_JobName    | <input checked="" type="checkbox"/> JOB_StartTime         |
| <input type="checkbox"/> Test2                 | <input type="checkbox"/> Test_restrict             | <input type="checkbox"/> Project                          |
| <input type="checkbox"/> Project_status        | <input checked="" type="checkbox"/> JOB_WorkDir    | <input checked="" type="checkbox"/> JOB_SubmissionUser    |
| <input checked="" type="checkbox"/> JOB_Status | <input checked="" type="checkbox"/> JOB_FinishTime | <input checked="" type="checkbox"/> JOB_SubmissionCommand |



JOB\_STATUS

- ☐ Empty value (9,965,670)
- ☐ 64 (31)

JOB\_WORKDIR

- ☐ Empty value (9,965,670)
- ☐ /gpfs/gpfs\_2mb/sgdemo/Power9 (31)

JOB\_FINISHTIME

- ☐ Empty value (9,965,670)
- ☐ 1558725413 (31)

JOB\_SUBMISSION...

- ☐ Empty value (9,965,670)
- ☐ sgdemo (31)

JOB\_SUBMISSION...

- ☐ Empty value (9,965,670)
- ☐ #BSUB -J gatk4;#BSUB -oo gatk4\_wex30x.err\_40c;%J;#BSU40;#;#;/gpfs/gpfs\_2mb/sgdemo/ (31)

JOB\_STARTTIME

- ☐ 1558724690 (31)
- ☐ Empty value (9,965,670)

JOB\_INPUTFILES

- ☐ Empty value (9,965,670)
- ☐ 0 (31)

JOB\_JOBNAME

- ☐ gatk4 (31)
- ☐ Empty value (9,965,670)

JOB\_JOBID



- ☐ Empty value (9,965,670)
- ☐ 673 (31)





# Spectrum LSF Provenance Data with Spectrum Discover


## Grouped search results



   Search

View results by: job\_submissionuser JOB\_JobId JOB\_JobName

Results:

Generate Report Add Tags Convert to individual record mode.



	job_submissionuser	job_jobid	job_jobname	Total Files	Total Size
	sgdemo	673	gatk4	31	3.87 GiB

Items per page: 20 | 1-1 of 1 items

1 of 1 pages < 1 >



# Spectrum LSF Provenance Data with Spectrum Discover



## View Individual Files Associated with workflow

								Add Tags
<input type="checkbox"/>	filename	size	job_workdir	job_jobid	job_submissionuser	job_jobname	job_status	job_starttime
<input type="checkbox"/>	time_bwa.log	1038.000	/gpfs/gpfs_2mb/sgdemo/Power9/wes_30x	673	sgdemo	gatk4	64	1558724690
<input type="checkbox"/>	gcat_set_025_bwa.bam	1979566750.000	/gpfs/gpfs_2mb/sgdemo/Power9/wes_30x	673	sgdemo	gatk4	64	1558724690
<input type="checkbox"/>	time_Markduplicates.log	860.000	/gpfs/gpfs_2mb/sgdemo/Power9/wes_30x	673	sgdemo	gatk4	64	1558724690



# Free 90-day

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.

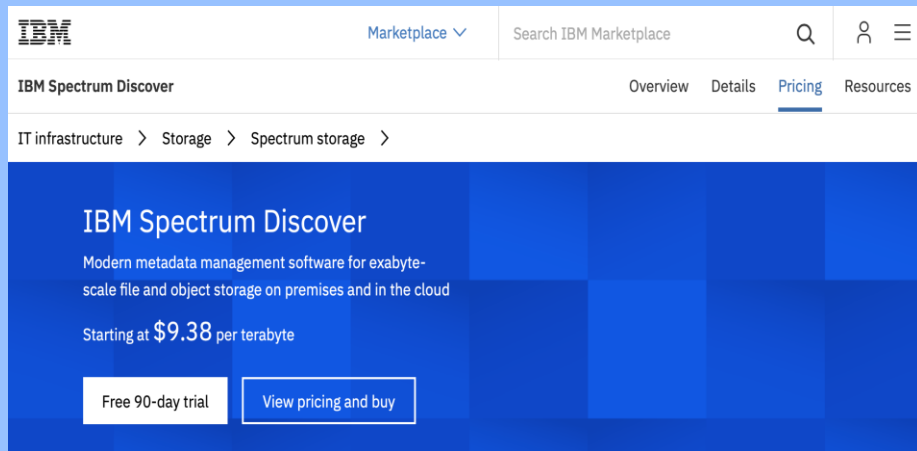
→ Free trial

[www.ibm.com/marketplace/spectrum-discover](https://www.ibm.com/marketplace/spectrum-discover)





# Learn more about Spectrum Discover



The screenshot shows the IBM Spectrum Discover product page on the IBM Marketplace. The page features the IBM logo in the top left, a search bar, and navigation links for Overview, Details, Pricing (which is highlighted), and Resources. Below the navigation bar, there is a breadcrumb trail: IT infrastructure > Storage > Spectrum storage >. The main content area has a blue background with a grid pattern. It displays the product name 'IBM Spectrum Discover', a description: 'Modern metadata management software for exabyte-scale file and object storage on premises and in the cloud', and the starting price: 'Starting at \$9.38 per terabyte'. At the bottom, there are two buttons: 'Free 90-day trial' and 'View pricing and buy'.

IBM Spectrum Discover

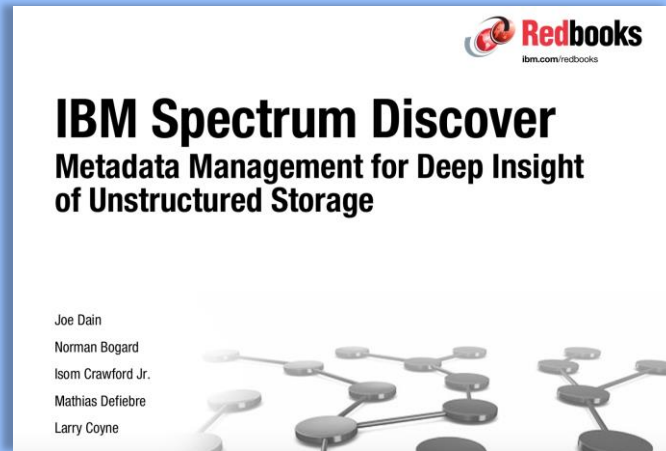
Modern metadata management software for exabyte-scale file and object storage on premises and in the cloud

Starting at \$9.38 per terabyte

Free 90-day trial View pricing and buy

Web Page and Customer Resources

[www.ibm.com/marketplace/spectrum-discover](http://www.ibm.com/marketplace/spectrum-discover)



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



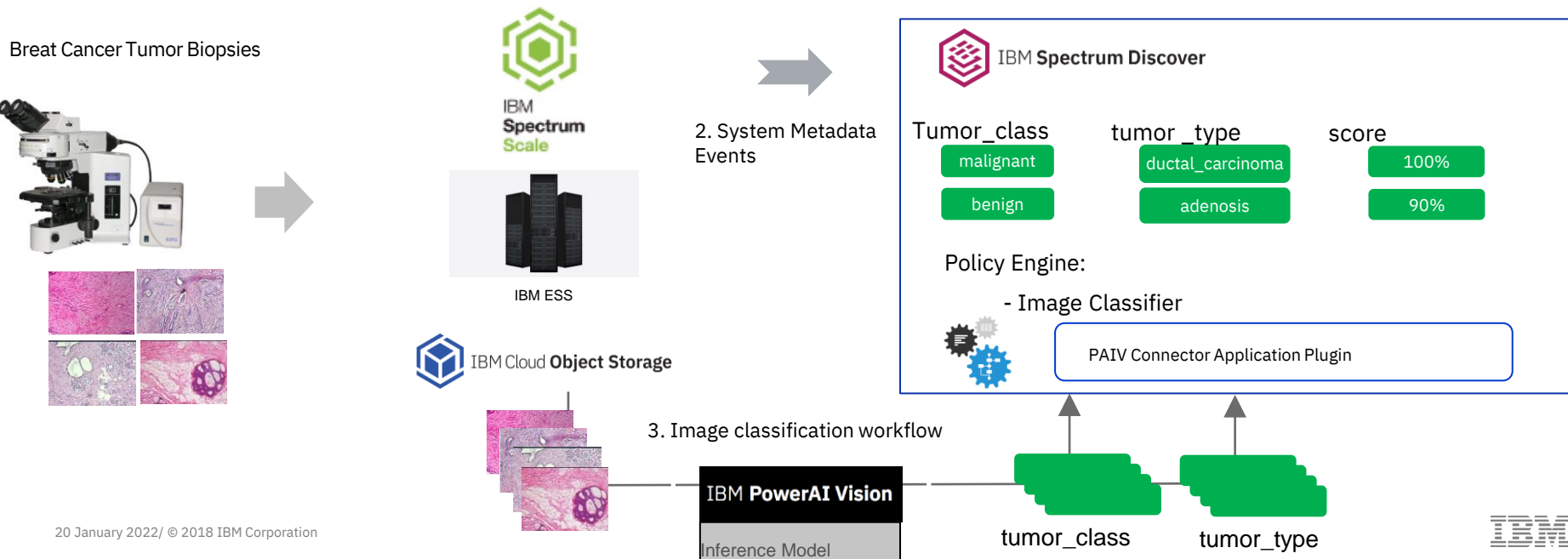
**IBM**



# 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 / COS

1. New imaging data ingested into Spectrum Scale / ESS, IBM COS storage
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 PowerAI Vision classification model, captures results and indexes into Spectrum Discover



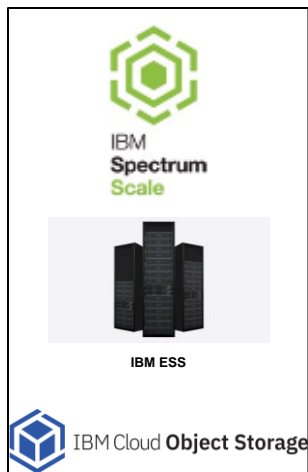
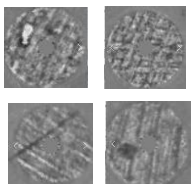


# Use Case: Automated Wafer Manufacturing Image Classification

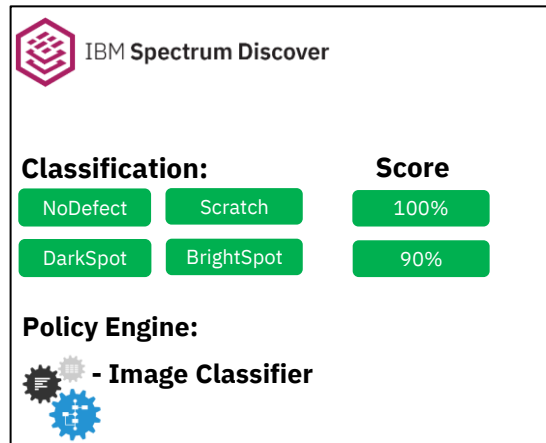
Event driven architecture to automatically classify and catalog wafer manufacturing data using PowerAI Vision inference model, Spectrum Discover, and Spectrum Scale / ESS / COS

1. New imaging data ingested into Spectrum Scale / ESS, IBM COS storage
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 PowerAI Vision classification model, captures results and indexes into Spectrum Discover

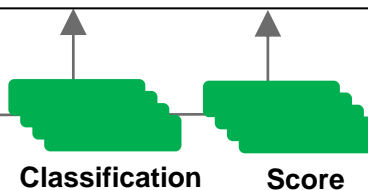
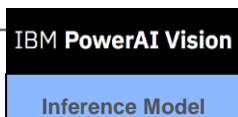
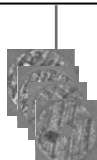
## 1. Wafer Manufacturing Imaging Data Ingestion



## 2. System Metadata Events



## 3. Image classification workflow





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

Data Mapping

The screenshot shows the 'Add new policy' form in a web application. The interface includes a sidebar with navigation links: Home, Search, Reports, Metadata, Admin, and Access. The main form is titled 'Add new policy' and contains several sections:

- Policy Type:** A dropdown menu set to 'CONTENT SEARCH'.
- Filter:** A text input field containing 'path like '/gpf/fs1/redbook/%''.
- Agent:** A dropdown menu set to 'contentsearchagent'.
- Tag:** A text input field containing 'containsPII'.
- Search Expression:** A list of checkboxes for selecting search expressions: 'Search Expression' (selected), 'EmailID', 'MasterCard', and 'US-SSN'.
- Value:** A dropdown menu set to 'True/False'.

Annotations with red callout boxes provide additional context:

- 'Select the CONTENT SEARCH agent' points to the 'contentsearchagent' dropdown.
- 'Specify that the data is not be shown, just the existence of PII' points to the 'True/False' value dropdown.
- 'Select all of the regular expressions to be used in the content search' points to the 'Search Expression', 'EmailID', 'MasterCard', and 'US-SSN' checkboxes.

At the bottom right, there are 'Save' and 'Cancel' buttons.

## 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

Generate reports

Drill down

The screenshot displays a web application interface for data discovery. At the top, a dark header bar contains a home icon, a search icon, and a 'Welcome sdadmin' message. Below the header is a sidebar with icons for Home, Search, Reports, Metadata, Admin, and Access. The main content area features a search bar with the query 'containspii in ('true') and path not like '/gpfs/fs1/redbook/governance/restricted/%'' and a 'Search' button. Below the search bar, there are two red callout boxes: 'Search for files that contain PII' and 'That are outside the designated PII storage location'. The search results are displayed in a table with columns 'path', 'filename', and 'owner'. The table lists six files, all owned by 'scooby'. To the right of the table is an 'Add Tags' button. A blue arrow points from the 'Generate reports' text to the 'Generate Report' button. Another blue arrow points from the 'Drill down' text to the 'Add Tags' button.

View results by:

Search for files that contain PII

That are outside the designated PII storage location

Results:

Generate Report

Add Tags

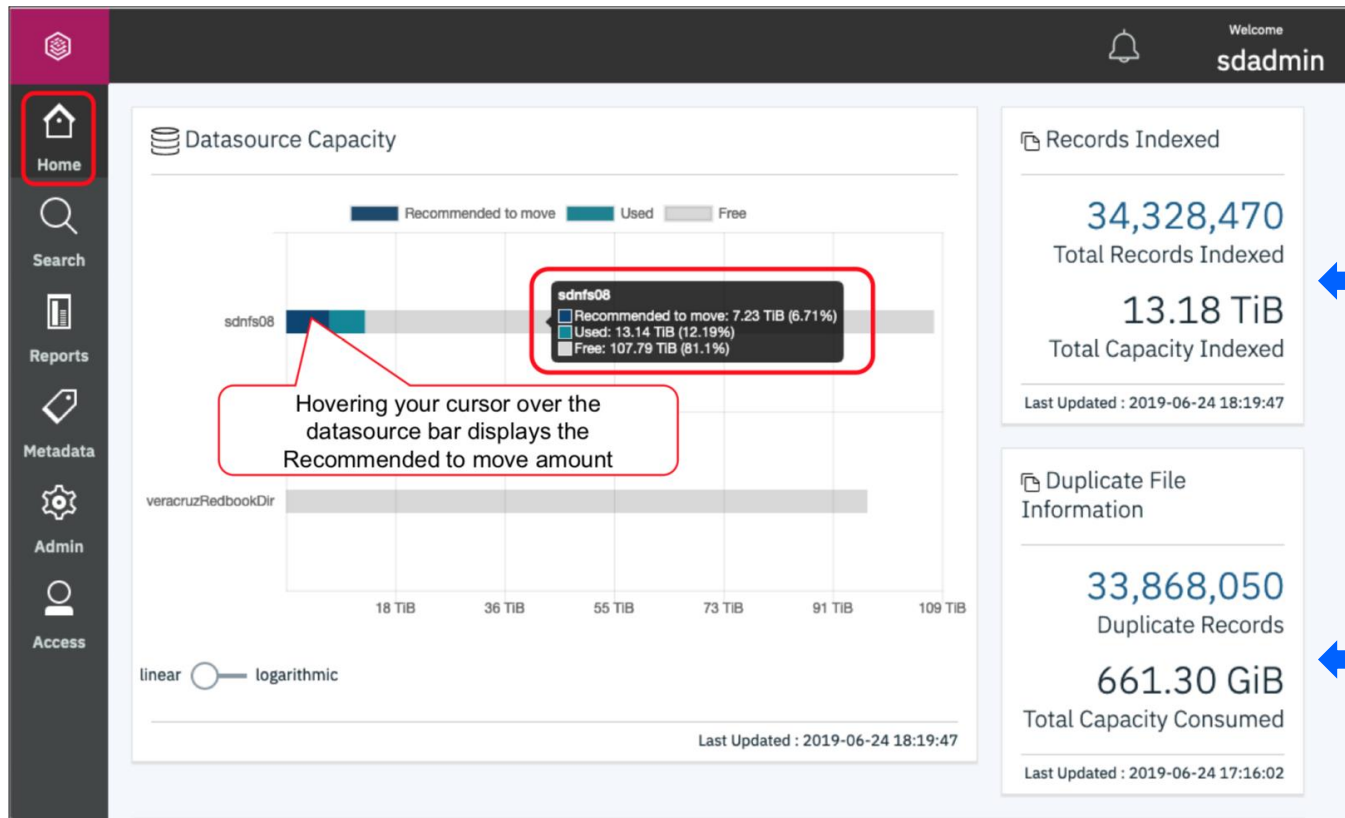
path	filename	owner
/gpfs/fs1/redbook/governance/shareall/eking/	roydmercer.dat	scooby
/gpfs/fs1/redbook/governance/shareall/eking/	other.csv	scooby
/gpfs/fs1/redbook/governance/shareall/eking/	mabell.info	scooby
/gpfs/fs1/redbook/governance/shareall/eking/	ccwindata3.csv	scooby
/gpfs/fs1/redbook/governance/shareall/eking/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

Customize view



# Discover in one screen duplicate records and data for archive

Data Visualization



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