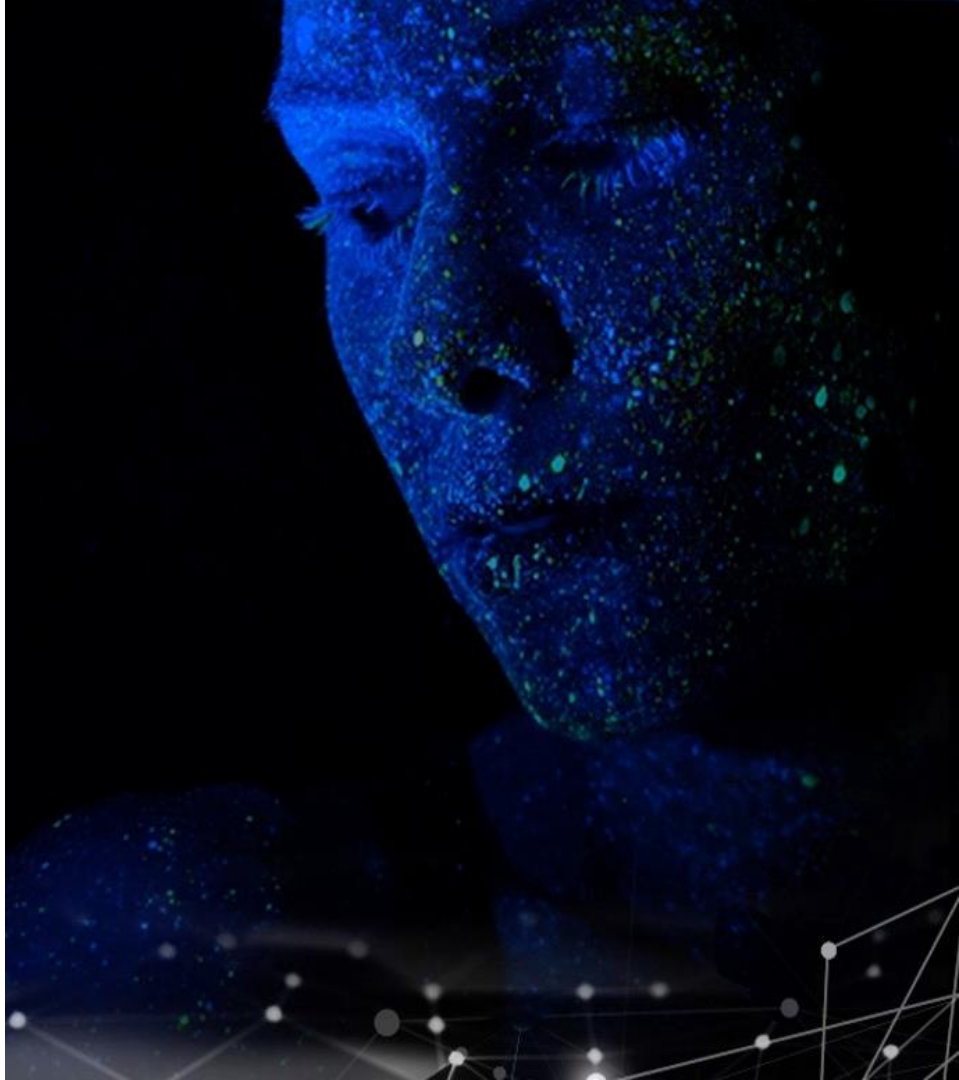
IBM Spectrum Discover

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

Indulis Bernsteins
Consulting Systems Architect
IBM UK

Agenda

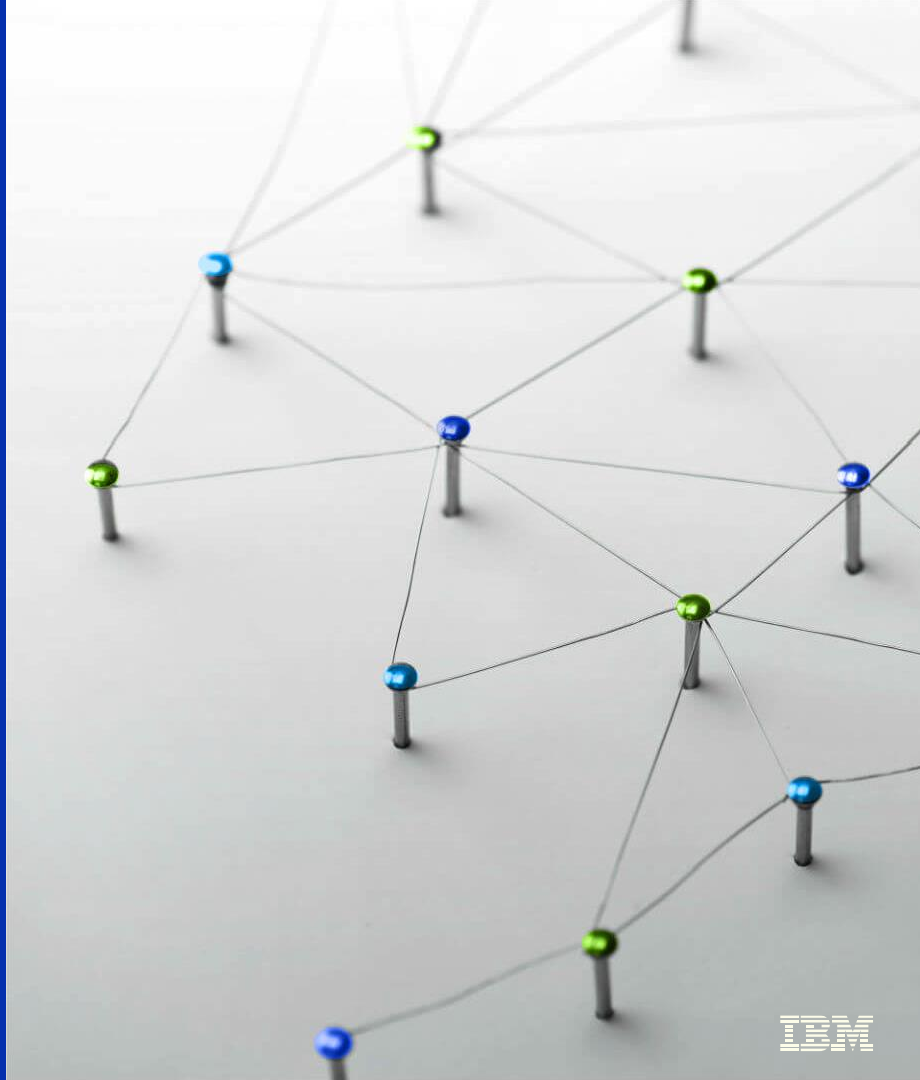
- Problem Statement
- Spectrum Discover Overview
- Use Case Discussion
- Demo



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

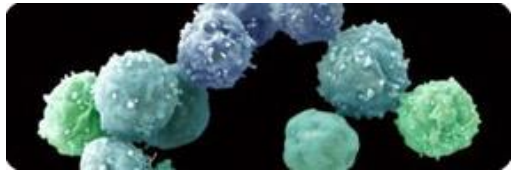
Metadata may be separate from the data, stored with the data, or derived from the data

- Posix inode plus extended attributes
- Standard document headers (doc, ppt, mp3, dicom, pdf, jpeg, GeoTIFF)
- Custom metadata tags
- AI derived metadata

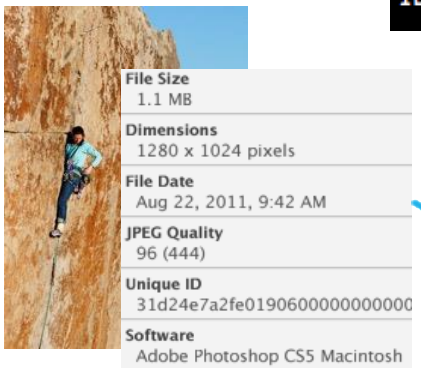
System Metadata

Location
Size
Owner
Group
Permissions
Last-Modified
...

Biomedical



Image



IBM PowerAI Vision

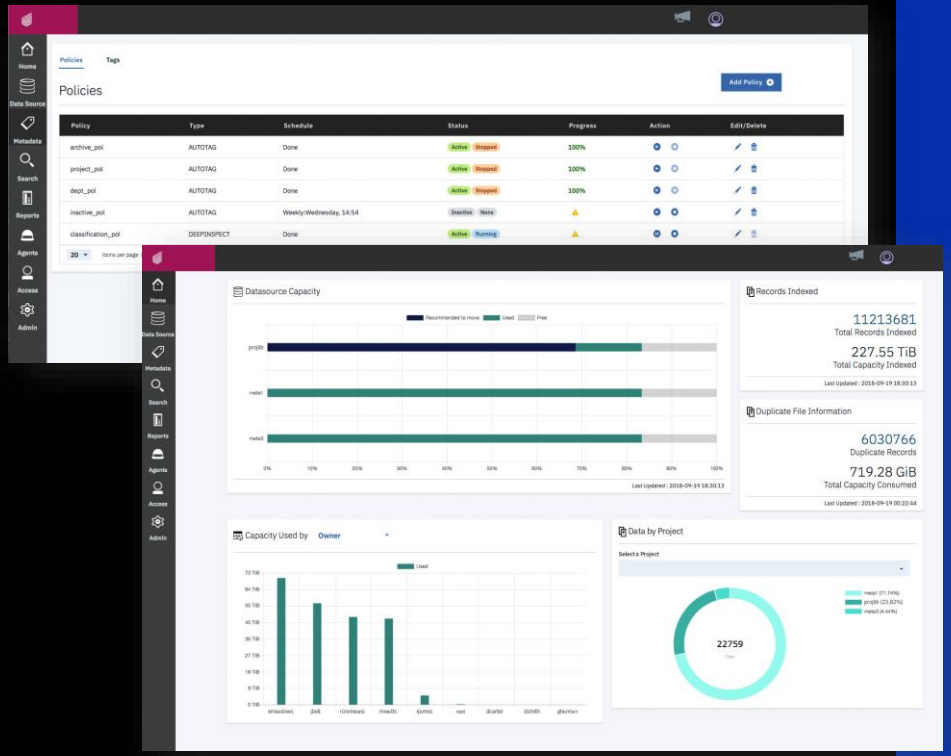
PYTORCH

TensorFlow

figure eight

Natural Language Processing

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

Data Insight

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

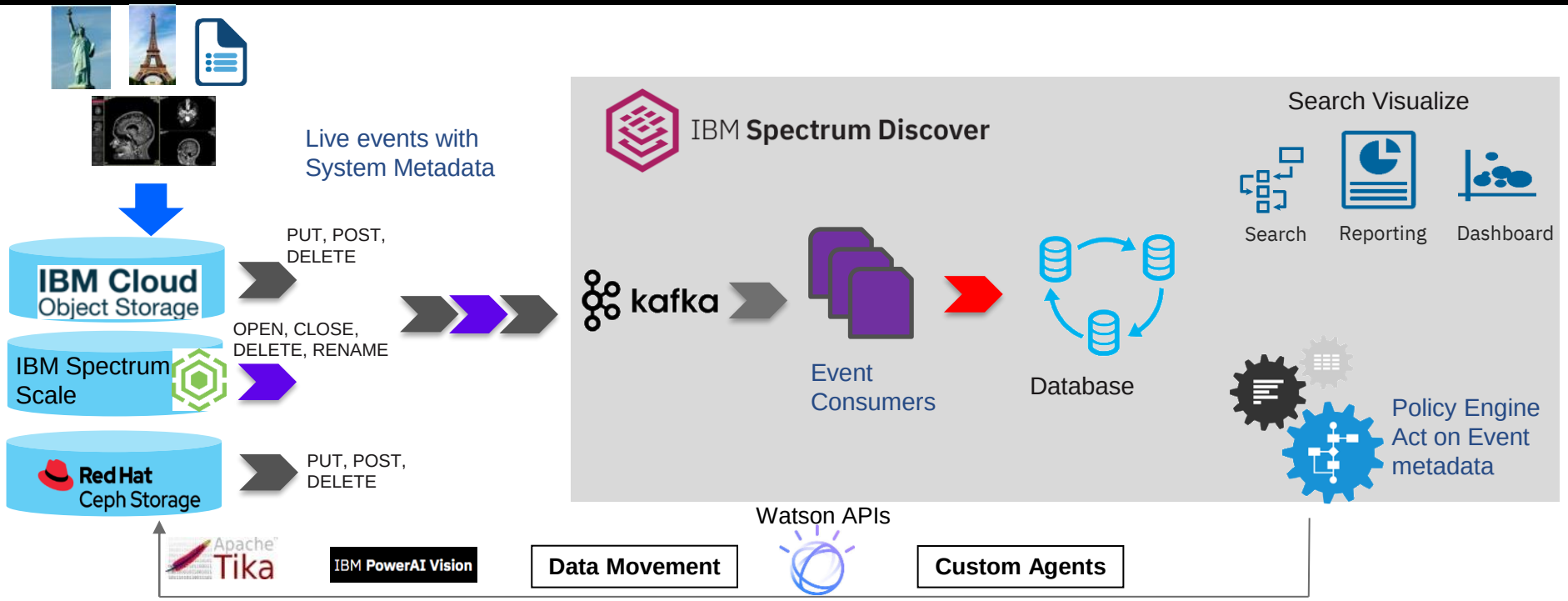
Activation & Optimization

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

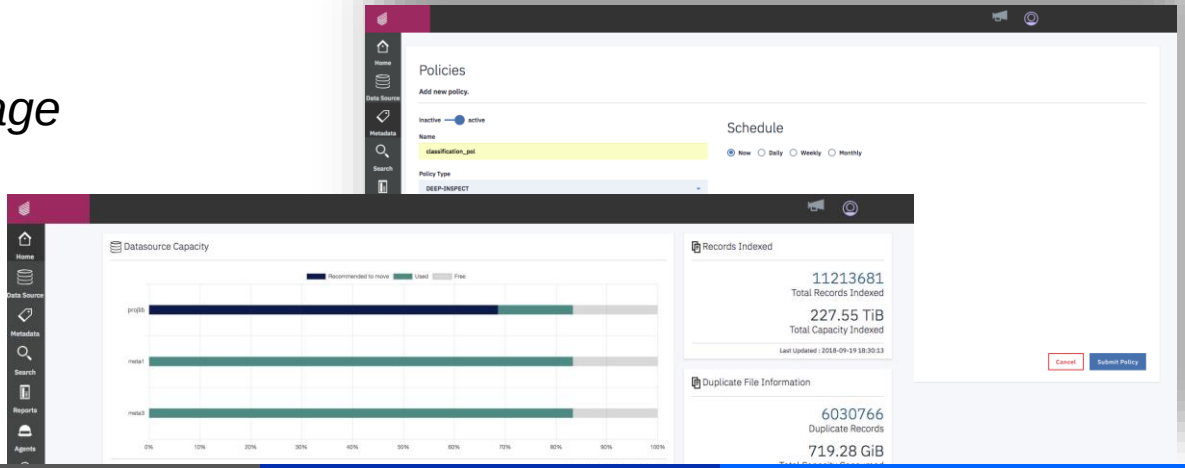


Harvest events from heterogeneous sources

- 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

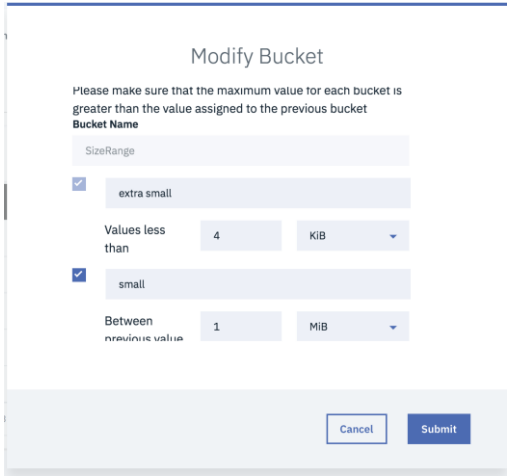
Actual results from PoC conducted with beta client in heterogeneous environment –
a major public health institution doing genomics research

File and Object Size Distribution Analytics

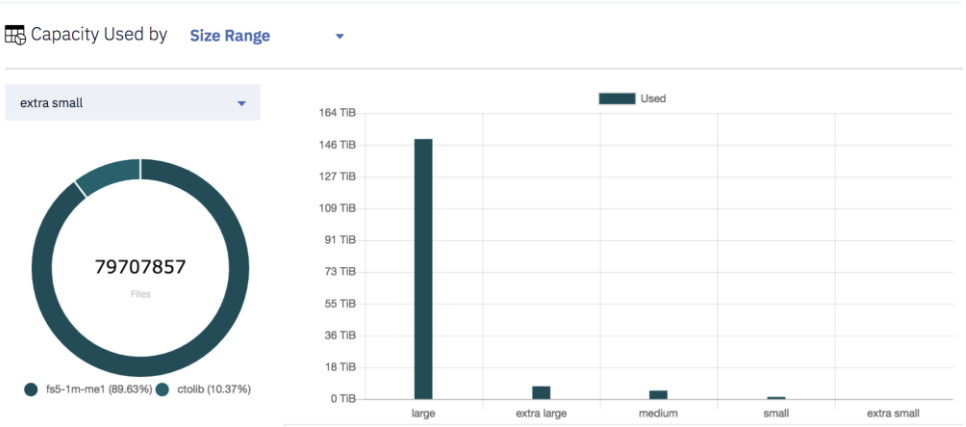
- Leverage size bucketing , visualization, and drill down search
- Default Bucketing

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

Customize Bucket Ranges

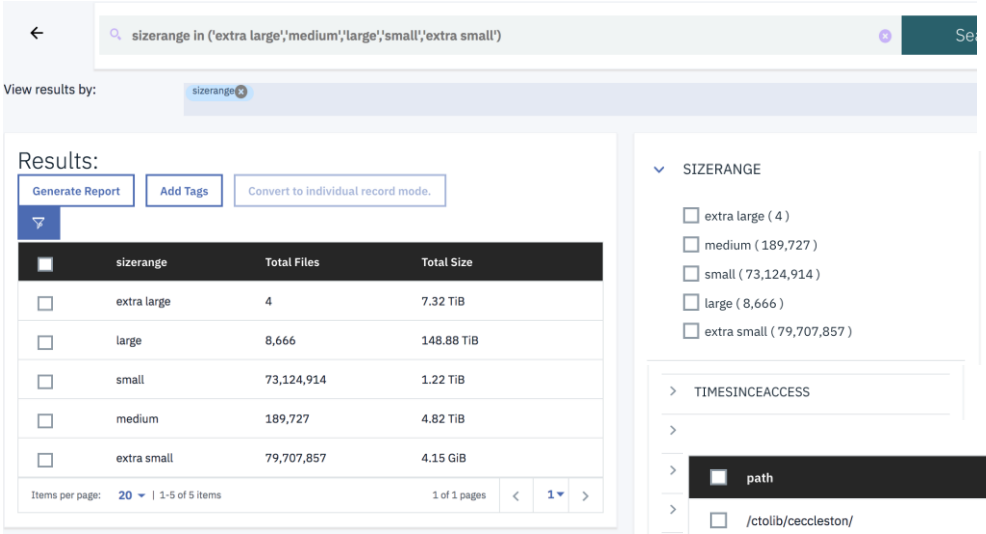


Visualize Capacity Usage by size bucketing

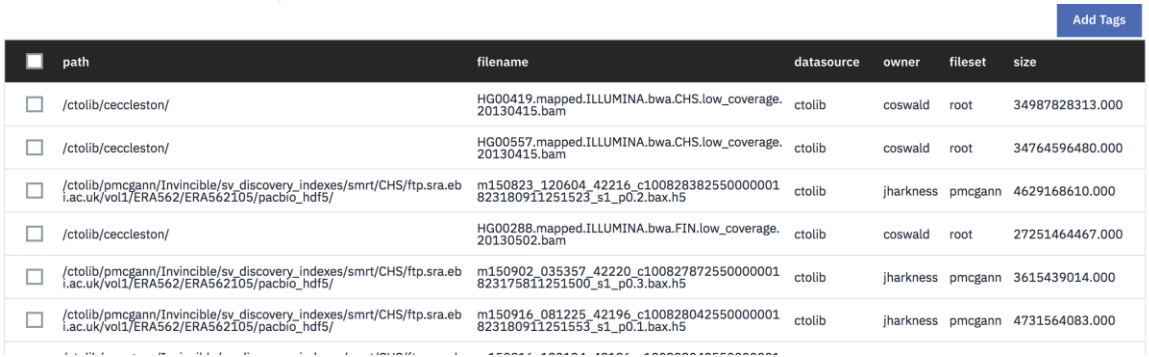


File and Object Size Distribution Analytics

Size range drilldown search



Convert to individual records

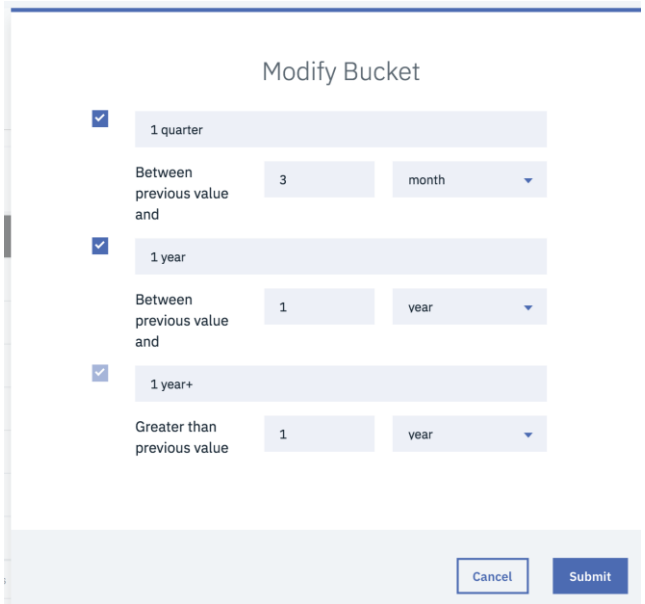


File and Object Data Aging Analytics

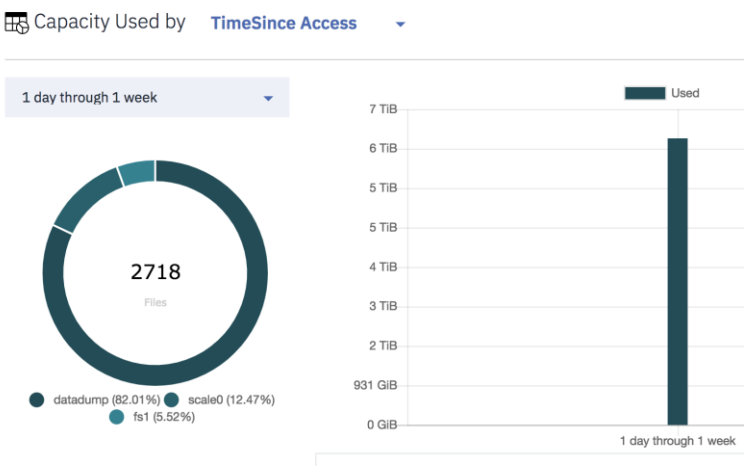
- Leverage time since access bucketing , visualization, and drill down search
- Default Bucketing

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

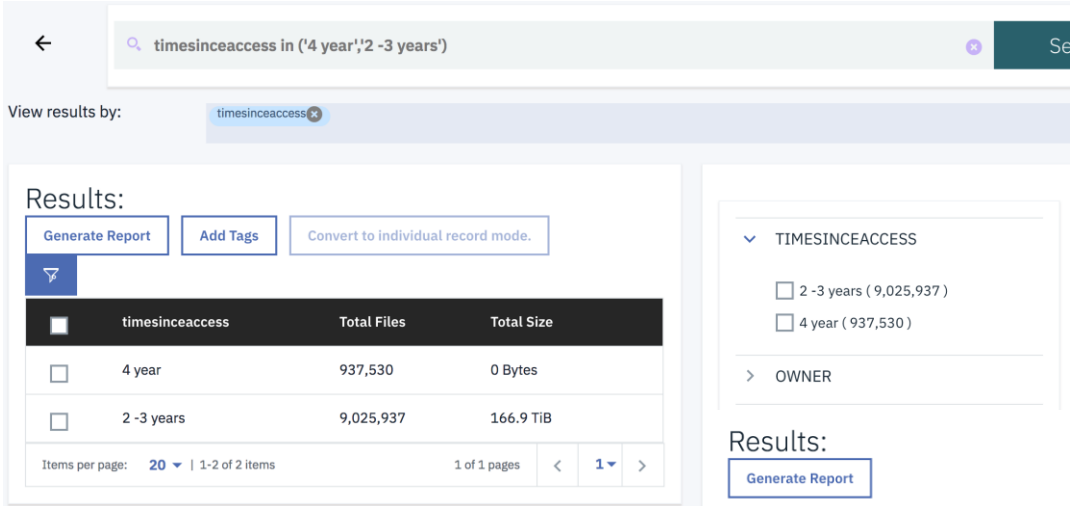


Visualize Capacity Usage by size bucketing



File and Object Size Distribution Analytics

Time range drilldown search



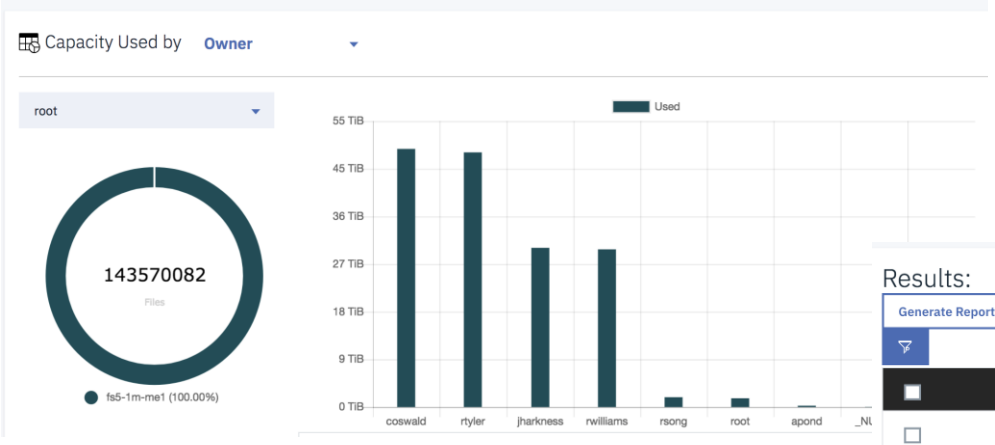
Convert to individual records

Add Tags

path	filename	datasource	owner	fileset	atime	size
/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
/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
/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
/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:

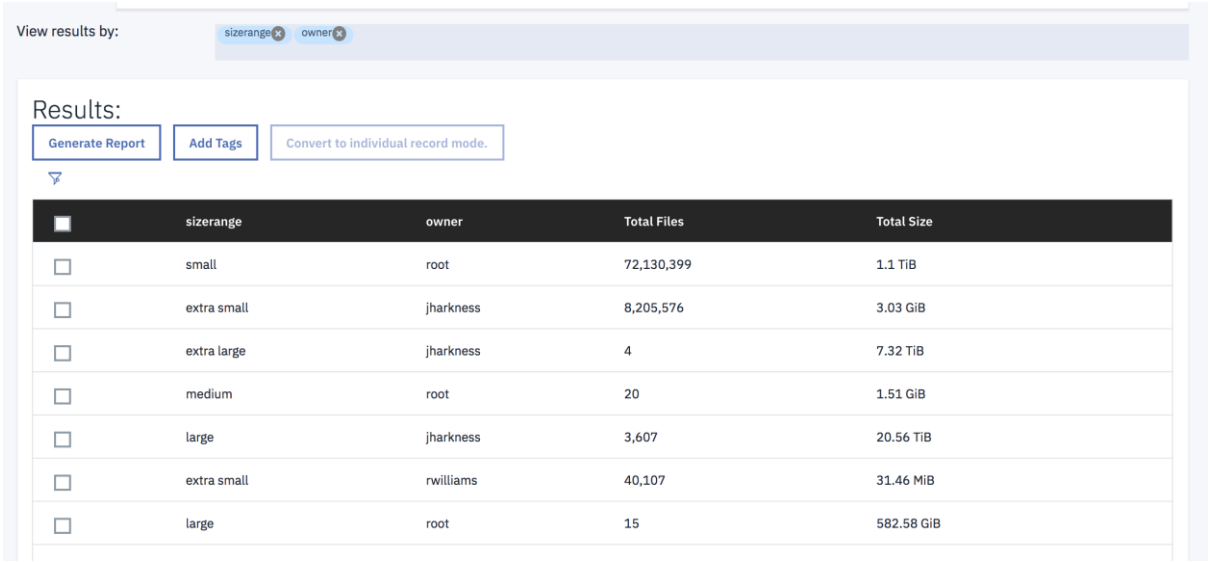
Generate Report | Add Tags | Convert to individual record mode.

owner	Total Files	Total Size
rtyler	417	48.61 TiB
jharkness	8,981,376	30.39 TiB
apond	506	268.54 GiB
rwilliams	104,035	30.1 TiB
NULL	521	54.46 GiB
root	143,570,082	1.67 TiB
coswald	4,426	49.27 TiB
nobody	2	7.82 KiB
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

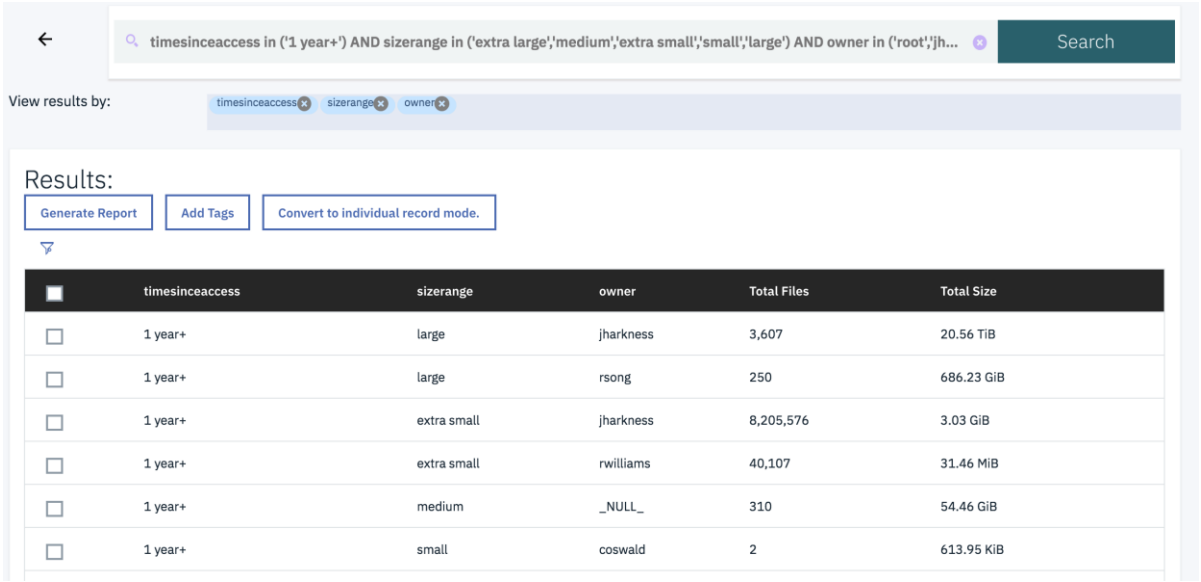
Example: space consumed by file size mapped against owner



sizerange	owner	Total Files	Total Size
small	root	72,130,399	1.1 TiB
extra small	jharkness	8,205,576	3.03 GiB
extra large	jharkness	4	7.32 TiB
medium	root	20	1.51 GiB
large	jharkness	3,607	20.56 TiB
extra small	rwilliams	40,107	31.46 MiB
large	root	15	582.58 GiB

Combine Criteria for Advanced Analytics

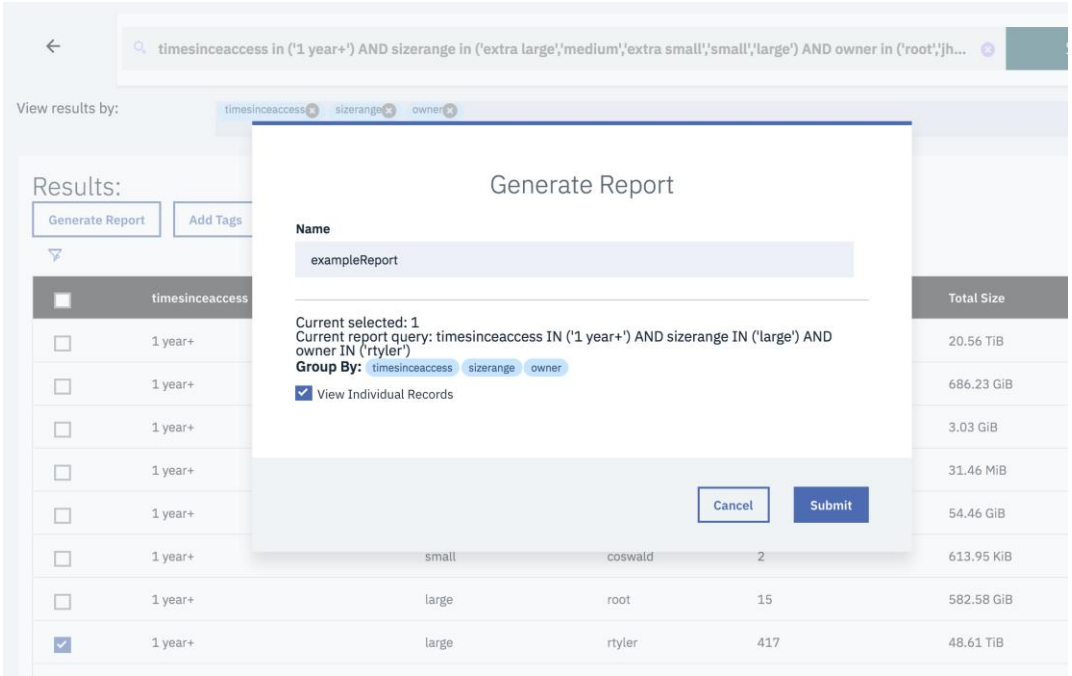
Example: space consumed by file size and time since access mapped against owner



timesinceaccess	sizerange	owner	Total Files	Total Size
1 year+	large	jharkness	3,607	20.56 TiB
1 year+	large	rsong	250	686.23 GiB
1 year+	extra small	jharkness	8,205,576	3.03 GiB
1 year+	extra small	rwilliams	40,107	31.46 MiB
1 year+	medium	_NULL_	310	54.46 GiB
1 year+	small	coswald	2	613.95 KiB

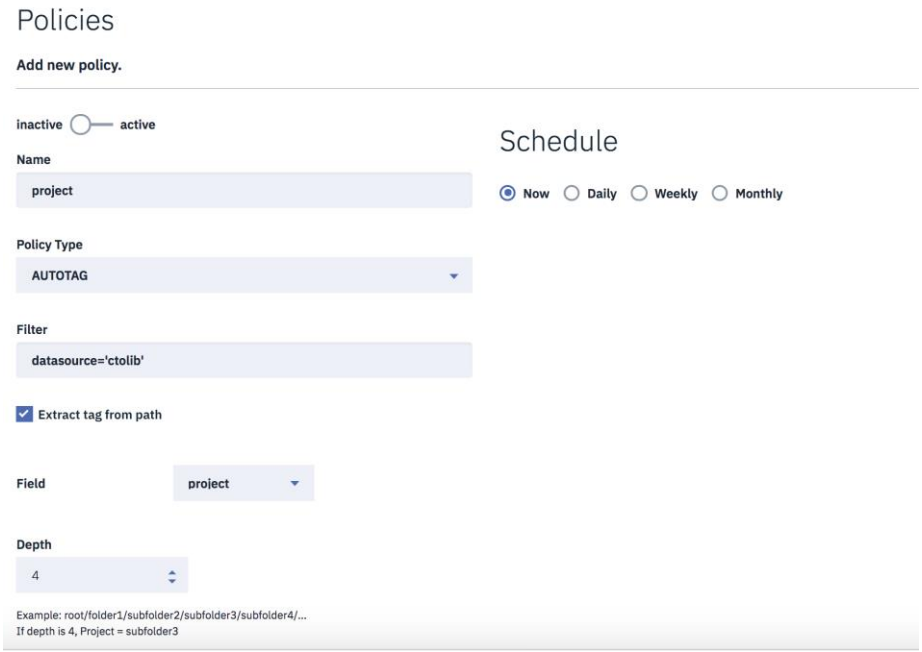
Generate Custom Reports

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



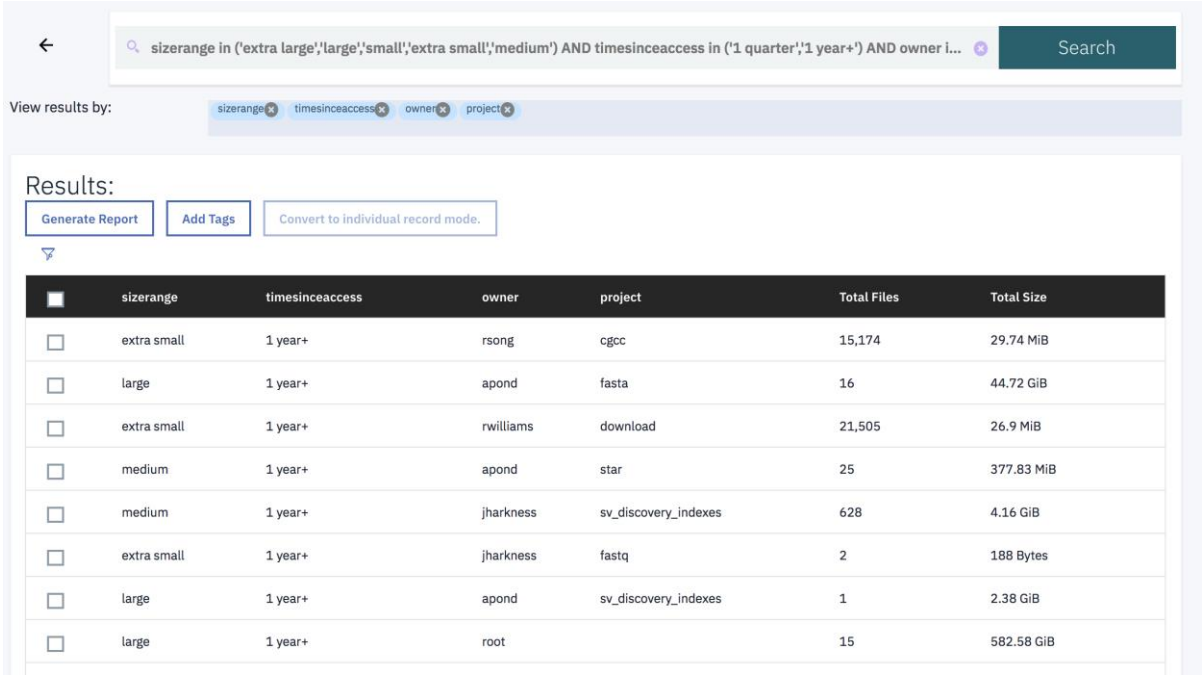
Map File and Object Data to Business Constructs

Example: Tag data by project



Map File and Object Data to Business Constructs

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

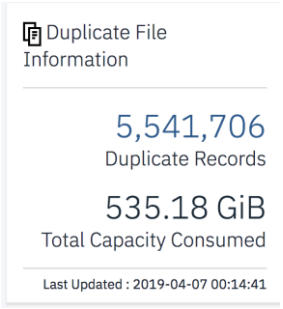


sizerange	timesinceaccess	owner	project	Total Files	Total Size
extra small	1 year+	rsong	cgcc	15,174	29.74 MiB
large	1 year+	apond	fasta	16	44.72 GiB
extra small	1 year+	rwilliams	download	21,505	26.9 MiB
medium	1 year+	apond	star	25	377.83 MiB
medium	1 year+	jharkness	sv_discovery_indexes	628	4.16 GiB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

Identify Potential Duplicate Data

Files with the same name and same size

Dashboard Analytics



Search Results

Results:

Generate Report | Add Tags | Convert to individual record mode.

filename	size	Total Files	Total Size
.dummy	0	2	
.local-guid	14	9,543	130.47 KiB
000001.out	436	9	3.83 KiB
000001.out	439	15	6.43 KiB
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

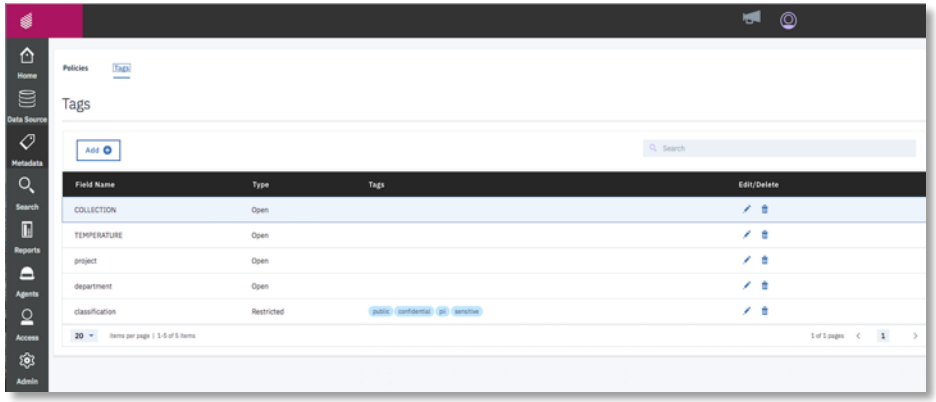
find PHI/PII from medical imaging datasets

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

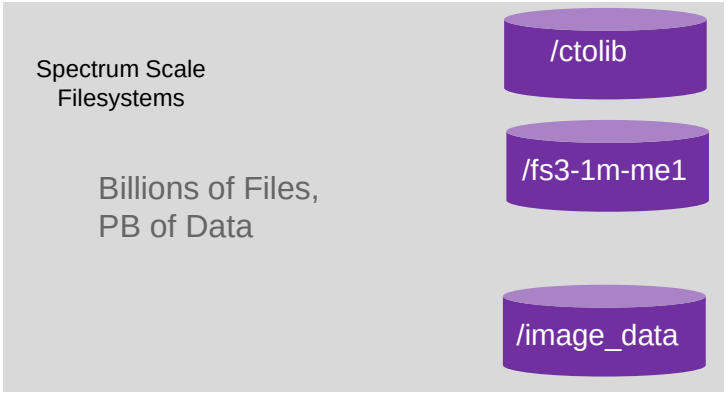
Metadata tags can be Open or Restricted

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

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

IBM Spectrum Discover

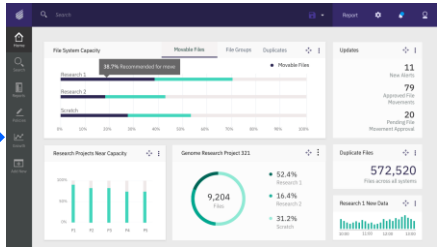
User	Department Tag	Project Tag	Project State Tag	Spectrum Scale Fileset / Base Directory
ibmuser1	staff	phase1	active	/whole_cell
ibmuser2	postdoctoral	phase2	inactive	/nucleus
ibmuser3		phase3	active	/polysomes



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

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

3. Ad-hoc filtered search



4. Generate reports (Capacity show back)

Capacity Reporting

5. Move to tape

IBM

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

IBM Spectrum Discover

Show me where my data resides for patents filed in 2016

IBM ESS

IBM Spectrum Scale

/ctolib

/fs3-1m-me1

.
.
.

/image_data

System Metadata Events

IBM Spectrum Discover

Tagsn:

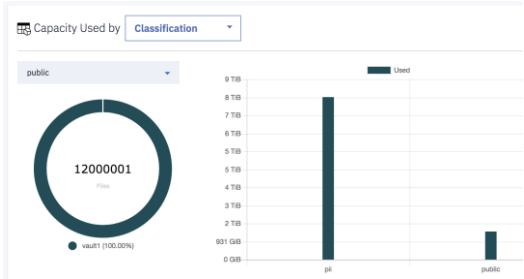
Patent_no

Patent_assignee

File_date

Policy Engine:

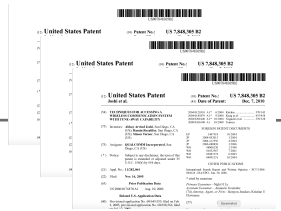
- Content Search



Header Extraction

VCF Format

PDF



IBM

Demo Use Case: Patent Inspection

IBM Spectrum Discover

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 Rezaiifar**, 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

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

IBM Spectrum Discover – Content Classification Workflow

IBM Spectrum Discover

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

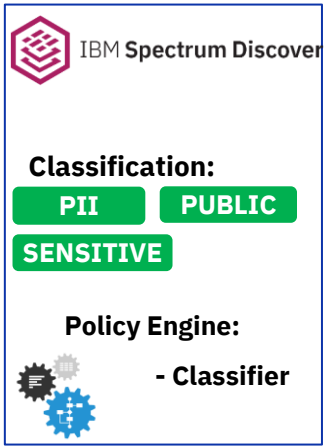
PII	Sensitive	Public
SSN	Confidential	Security: None
DOB		
Phone #		

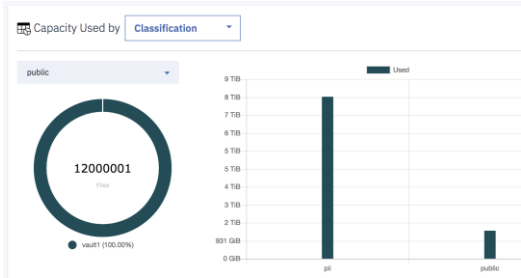
What kind of sensitive data is stored in my enterprise?

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

Where is this data stored?

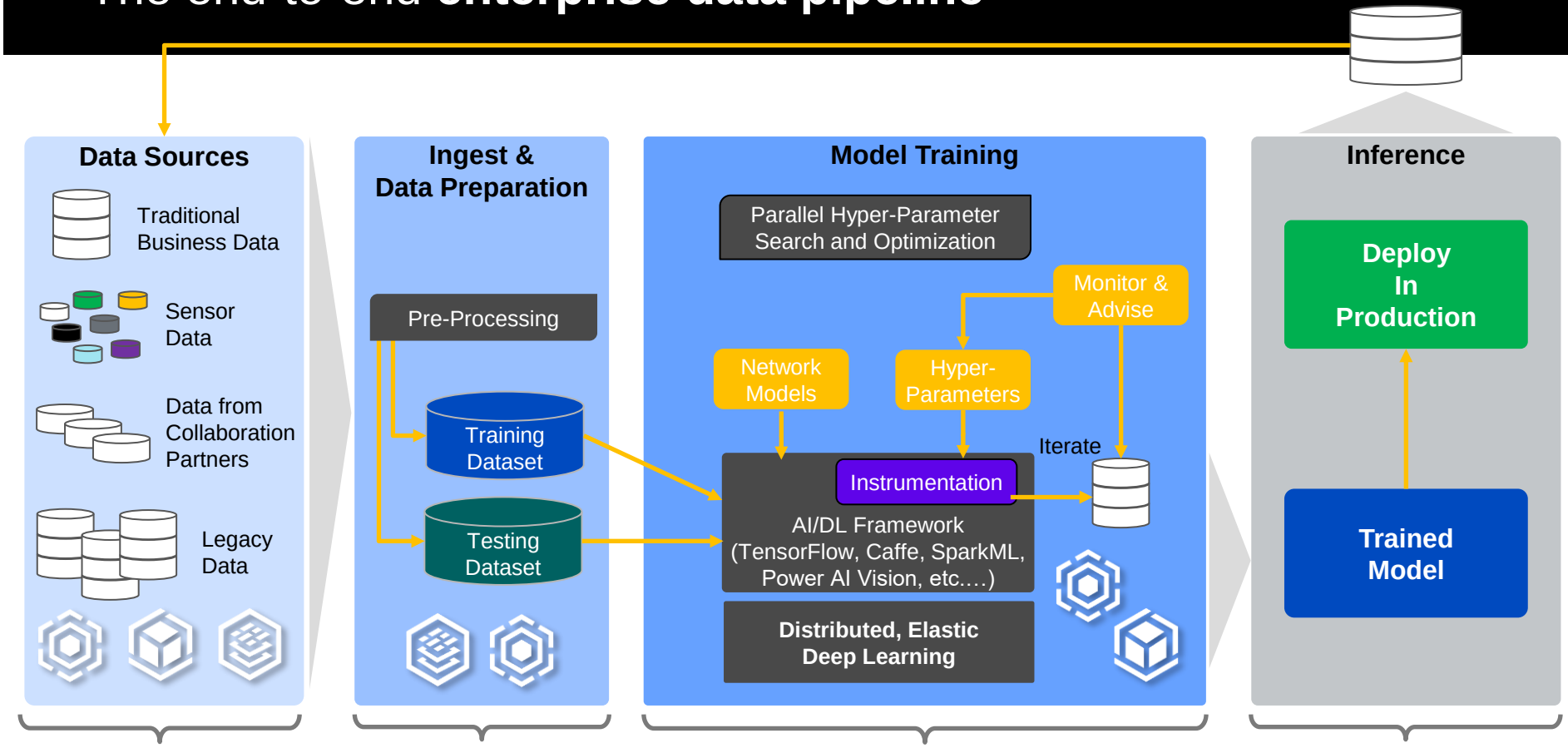
System metadata





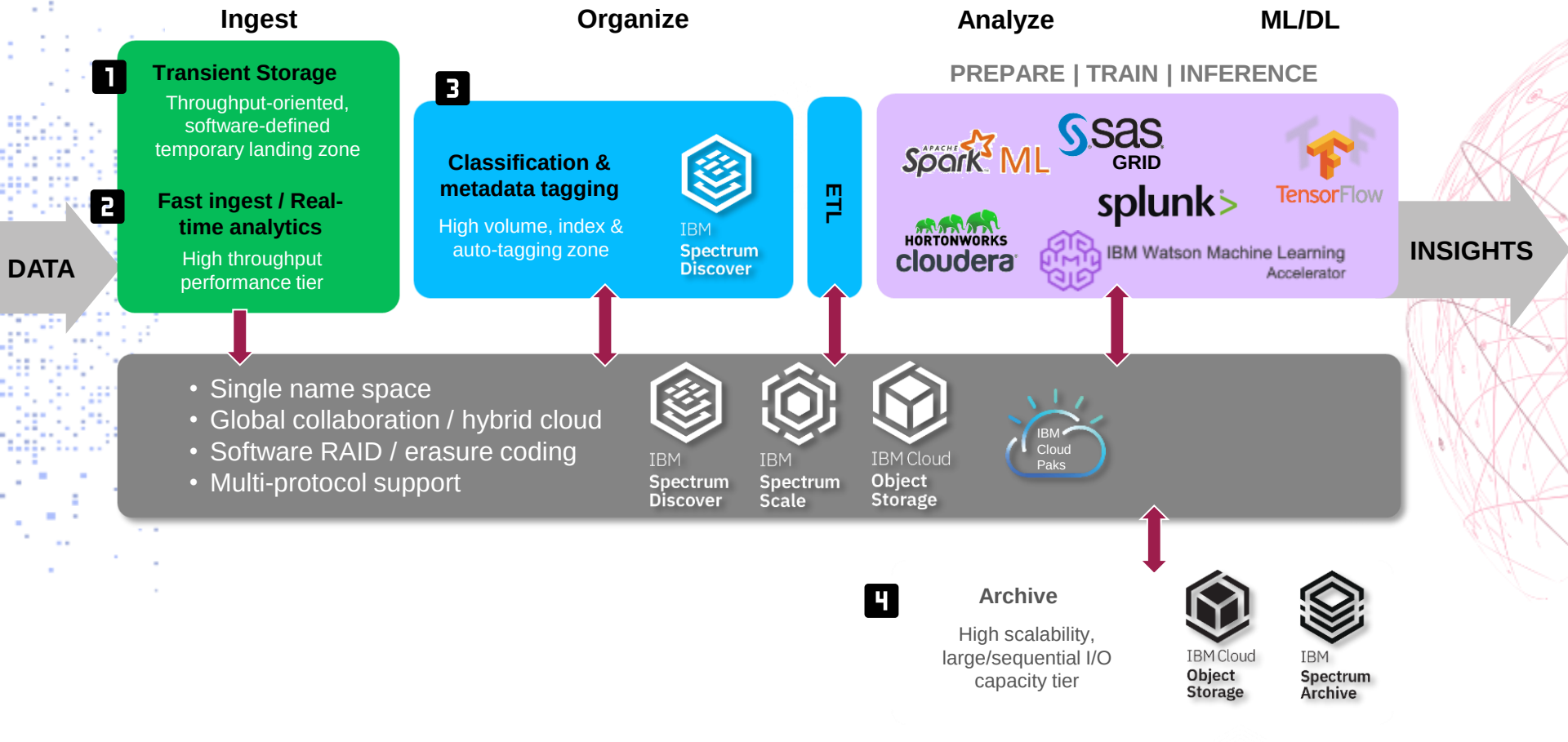
Spectrum Discover and the AI Pipeline

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.

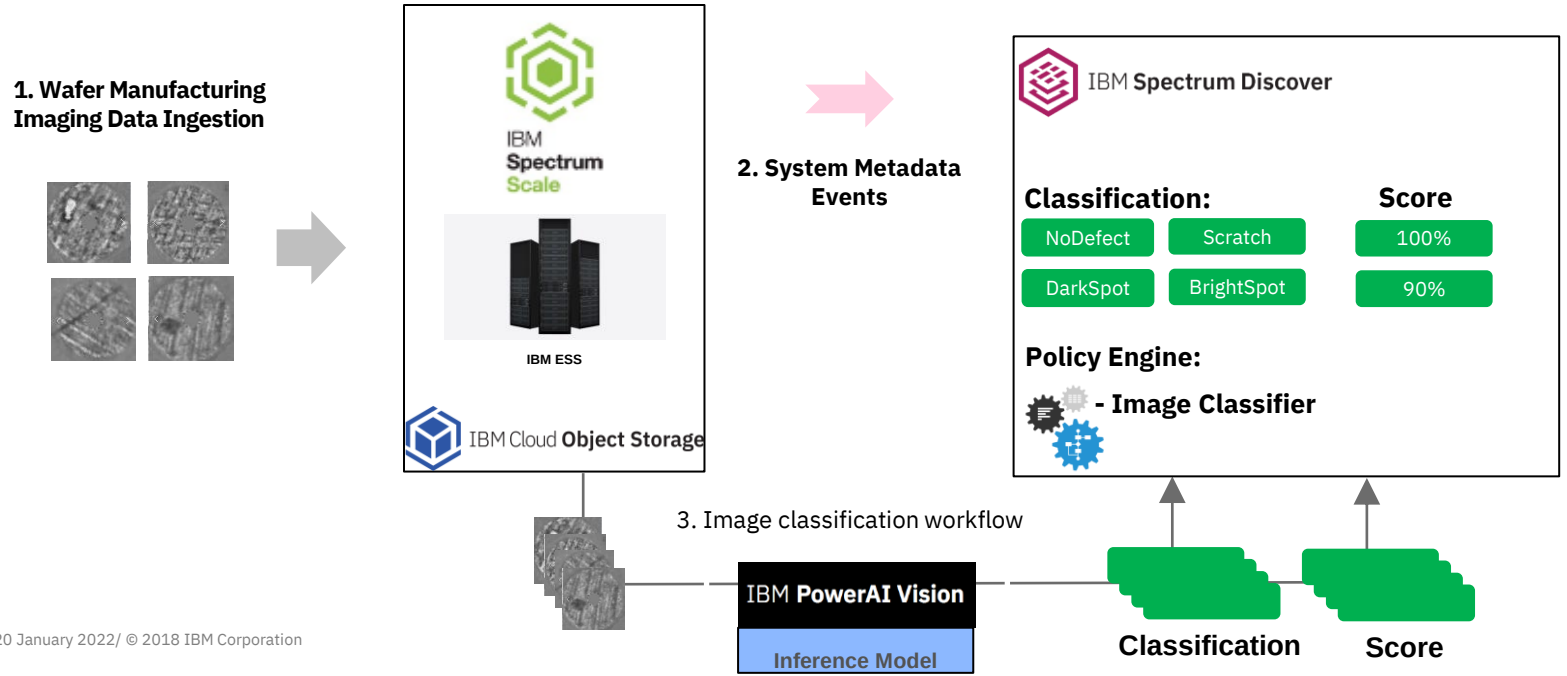
slack

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.

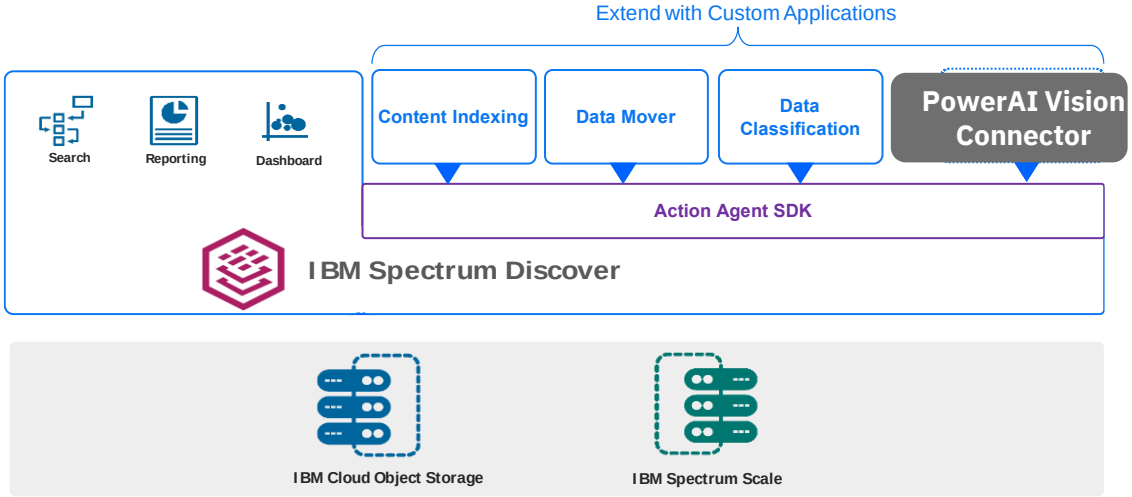
Use Case: Automated Wafer Manufacturing Image Classification

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



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

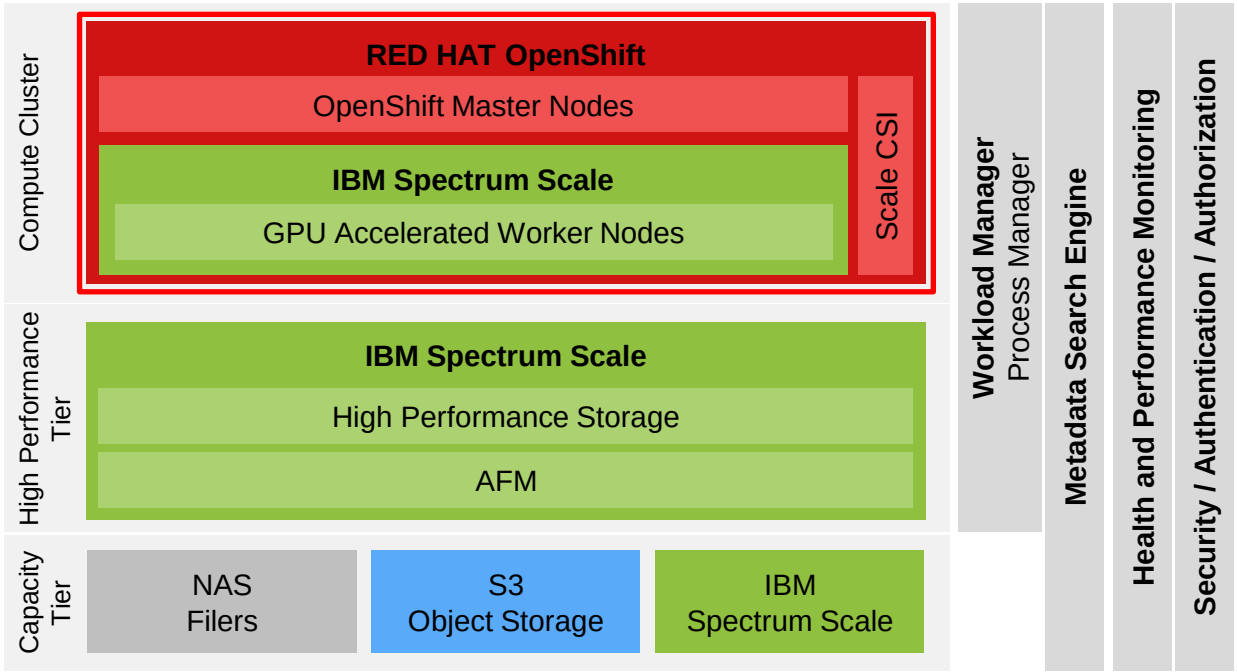
GitHub

docker hub

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

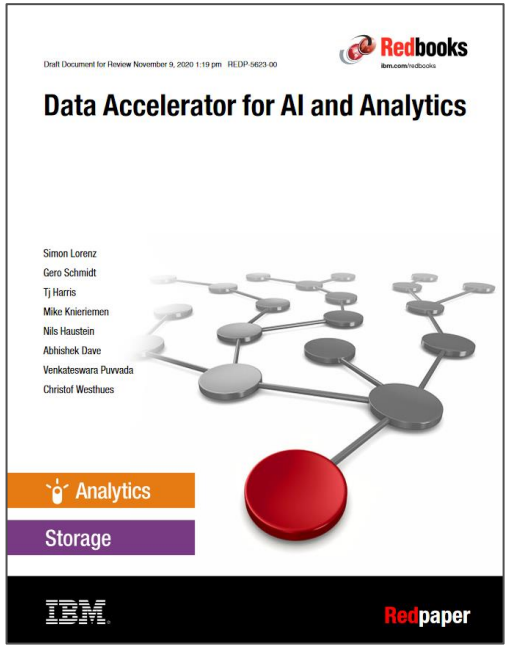
Data Accelerator for AI and Analytics

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)



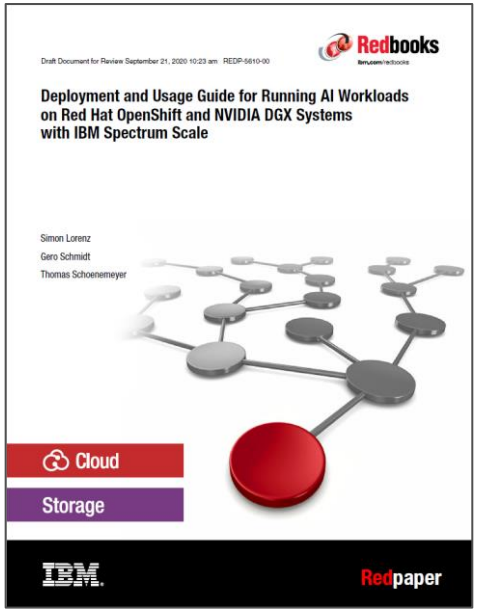
AI Example Use Case: Autonomous Driving

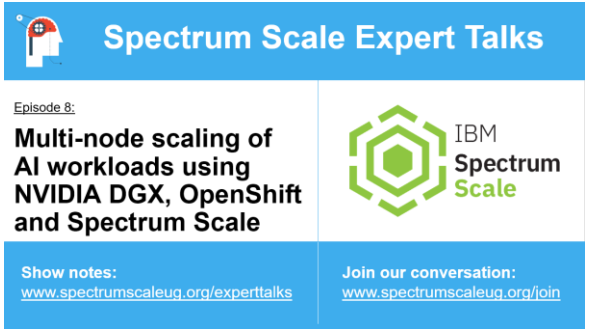
Worked on an IBM Redpaper:

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





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

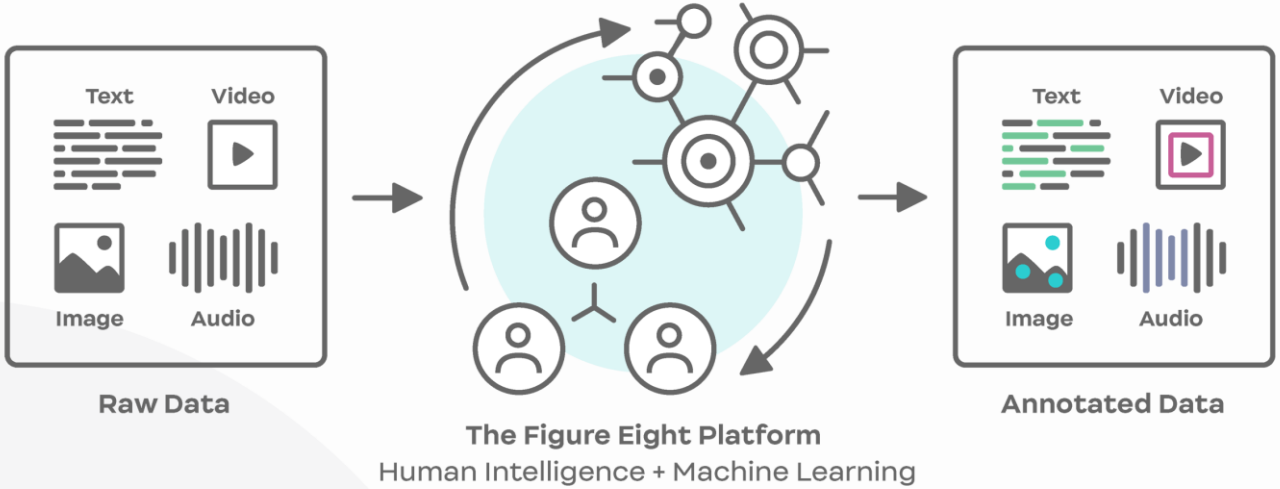
How is ai helping fire fighting?

- 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

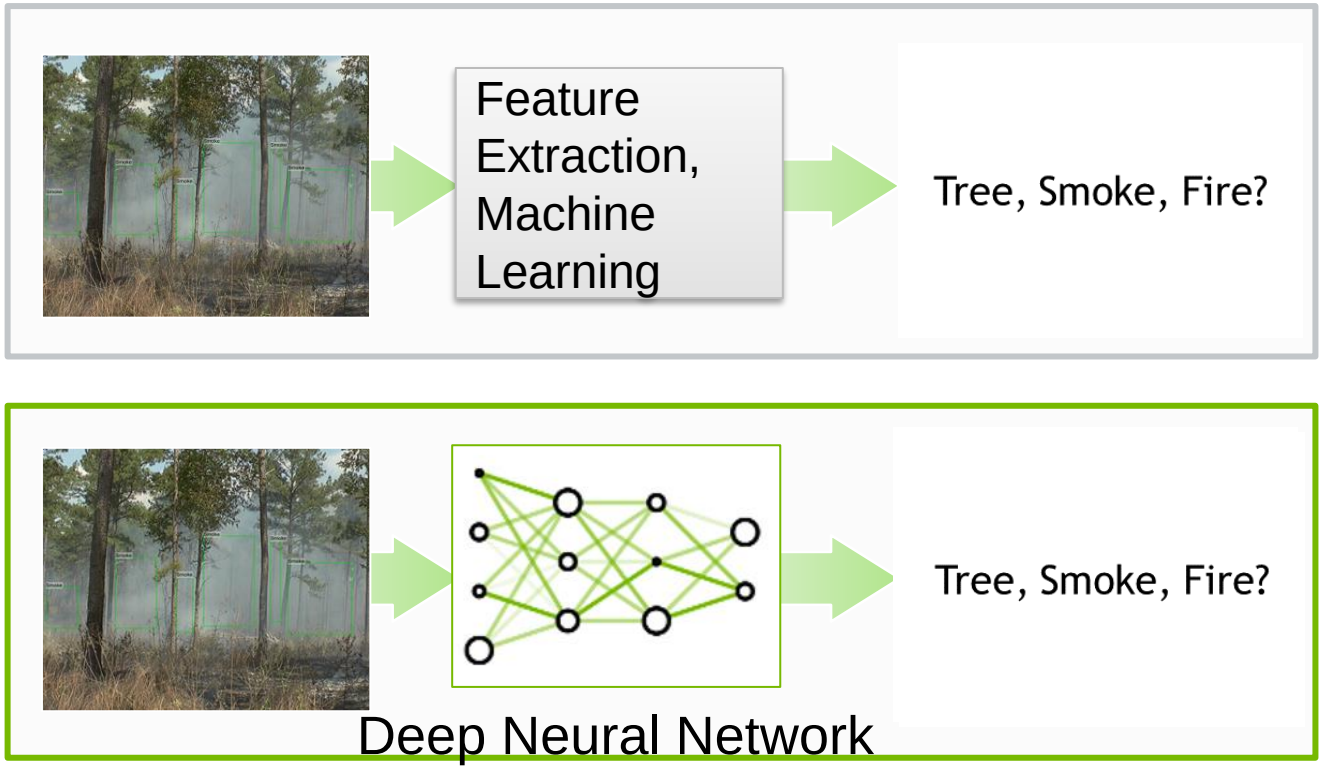
figure eight



If you need labels and annotations for your machine learning project, we can help. You upload your unlabeled data, with the rules you need for your machine learning project, and launch. We use a distributed network of human annotators and cutting edge machine learning models to annotate that data at enterprise scale.

IBM Wildfire Dataset

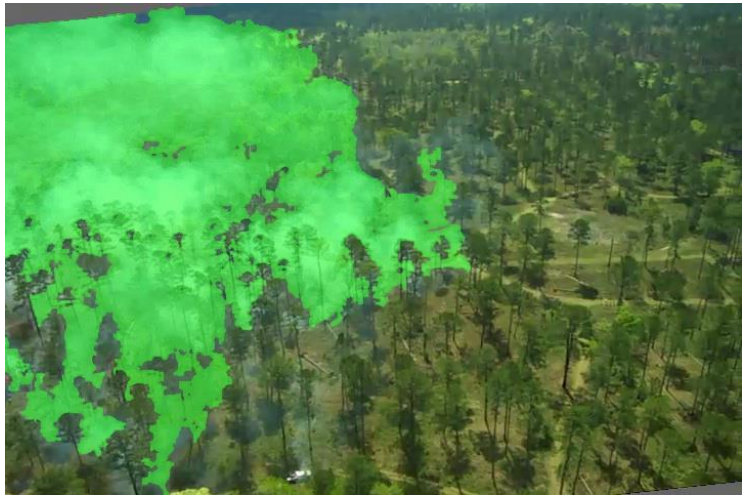
Identify smoke at fire line for prescribed burns



How did Figure-eight annotate the wildfire dataset

Semantic Segmentation

Pixel-level labeling for computer vision projects



Ontology

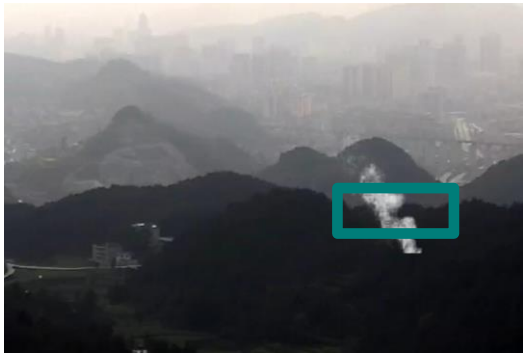
display_color	description	class_name	output_value
#21ff17		Smoke	1

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	12.248.233.9	{"url":"https	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	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	12.248.233.9	{"url":"https	FALSE		http://nickfigure8.s3.amazonaws.com/Nvidia%20pics/video_261_0047989.jpg

How did Figure-eight annotate the wildfire dataset

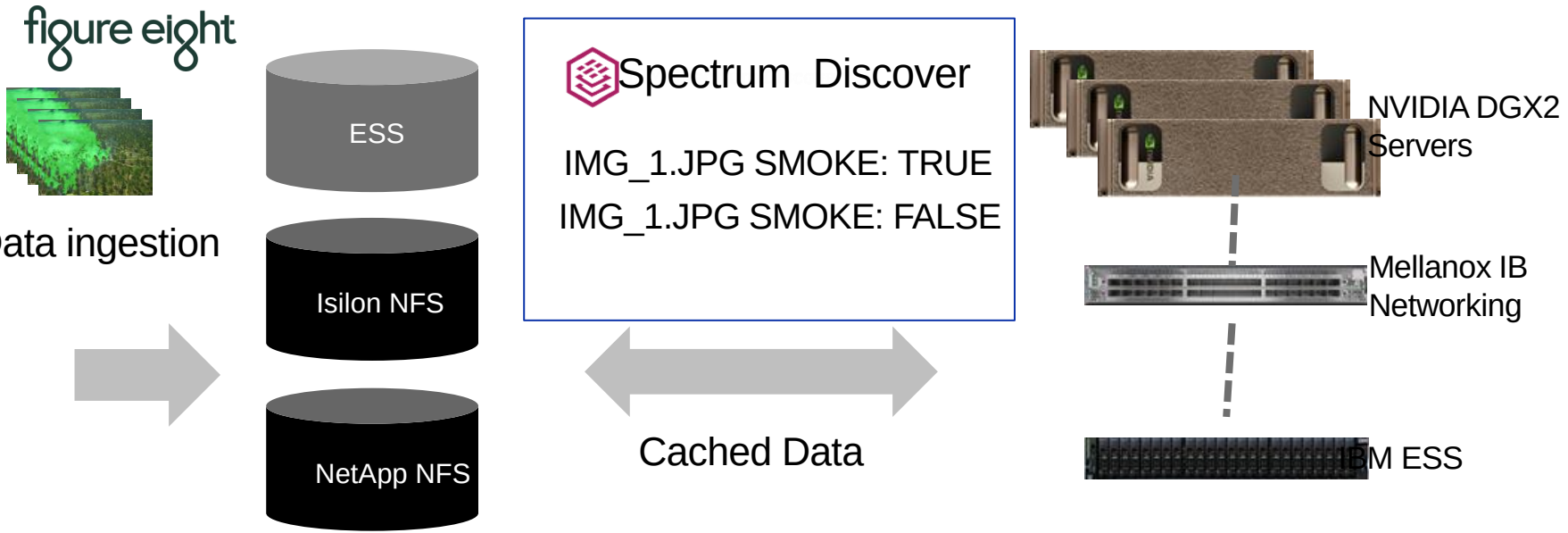
Bounding Box Object Detection
Polygon based bounding box annotations on wildfire dataset



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

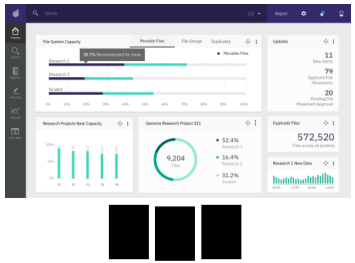
File caching/prefetching w/Spectrum Discover leveraging Figure-eight annotations

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



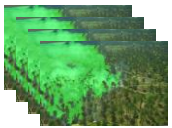
Filtered Caching with Spectrum Discover Based on Labels / Annotations

Cache images from San Bruno that have smoke in them



Datasource	Path	Filename	Country	Region	City	Category	Visbility
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



ESS

Spectrum Discover

Cached Data:
video_261_0047987.jpg

NVIDIA DGX2 Servers

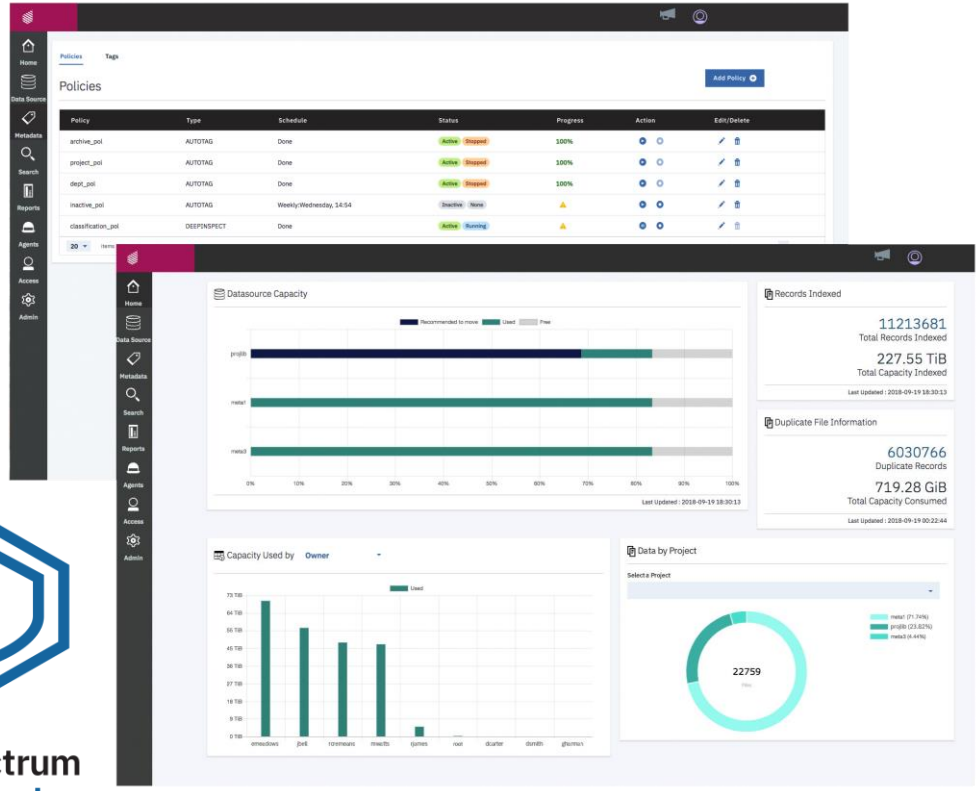
Mellanox IB Networking

IBM ESS

Spectrum Discover Support for Spectrum Protect

Gain deep insights into data in backup environments

with support for IBM Spectrum Protect



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

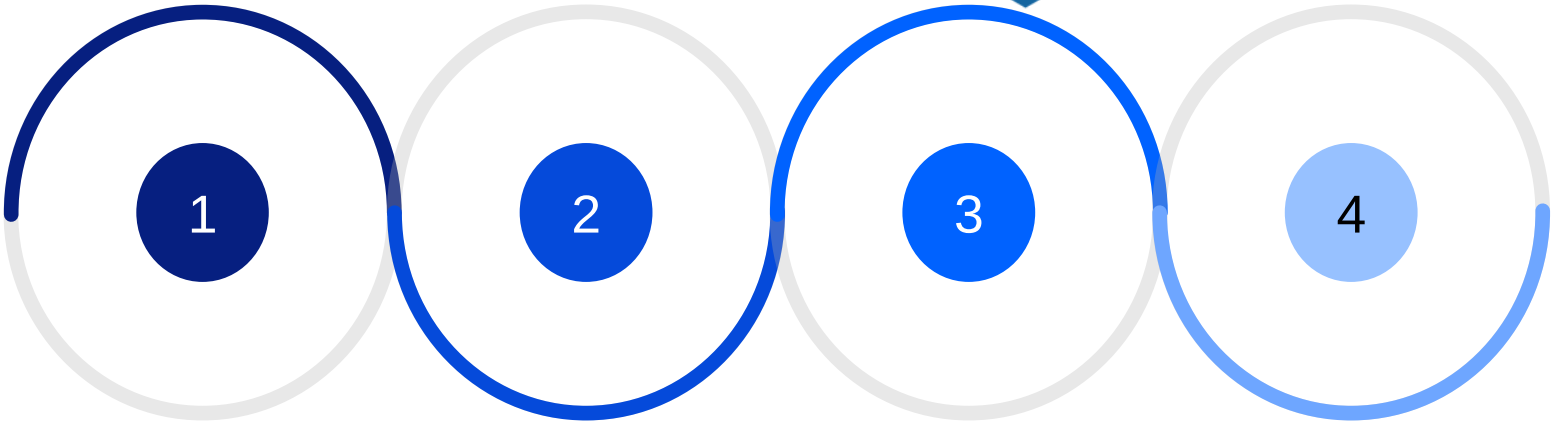
5 Common data protection questions

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

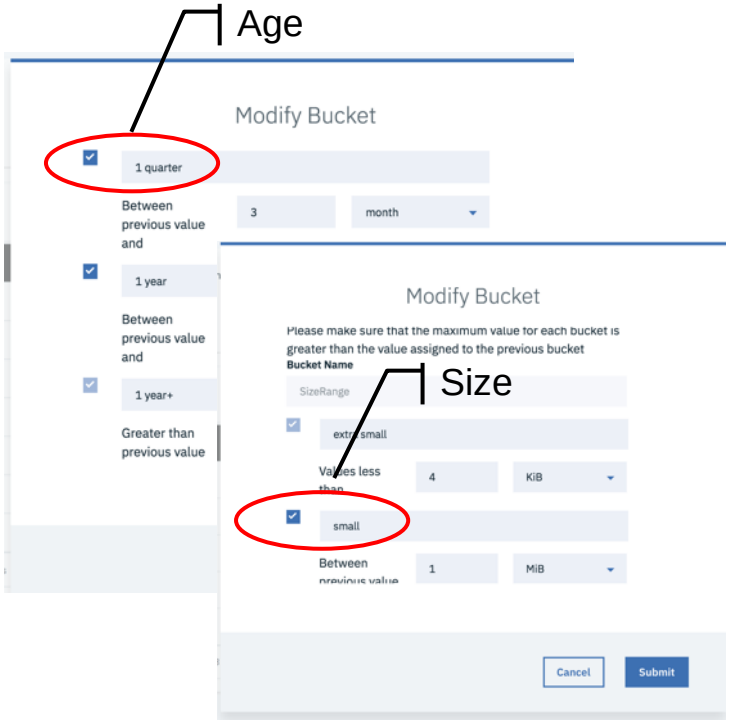
IBM **Spectrum Discover** + IBM **Spectrum Protect**



1. Ingest system metadata from Spectrum Protect
2. Utilize default analytics and generate reports
3. Customize analytics and reports
4. Map analytics against one or more system metadata attributes

Find | Curate and organize | Inspect and classify | Optimize

“Bucketize your data



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

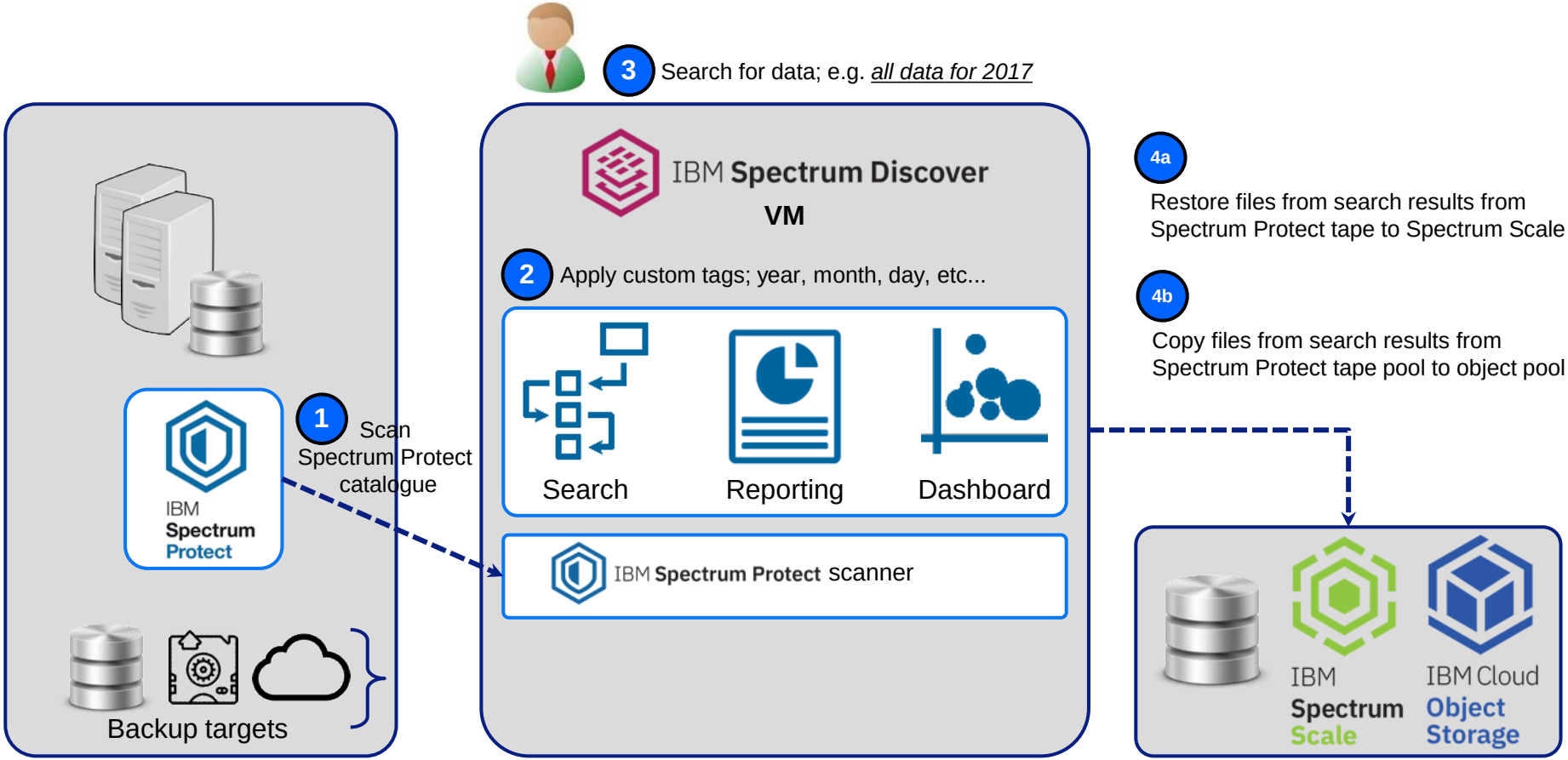
Select specific criteria for further analysis

Example – Ad hoc search

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?z

Transformations

- What algorithms were used?
- What transformations were applied?
- What parameters were used?
- How many iterations?
- What it cleaned, filtered, altered?
- Do we have intermediate results?

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

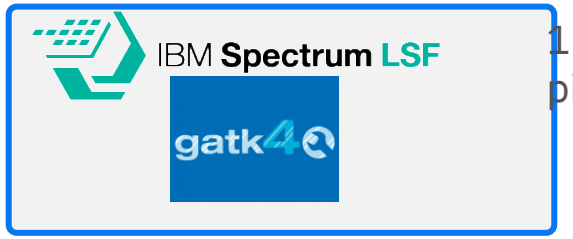
Manual cataloging is
- Inconsistent
- Incomplete
- Lacks formalism
- Cross group collaboration is prohibitively cumbersome

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

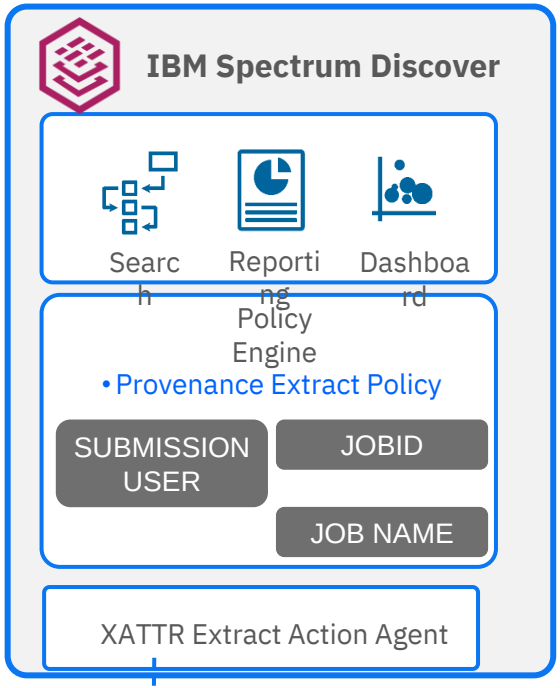
IBM Spectrum Discover



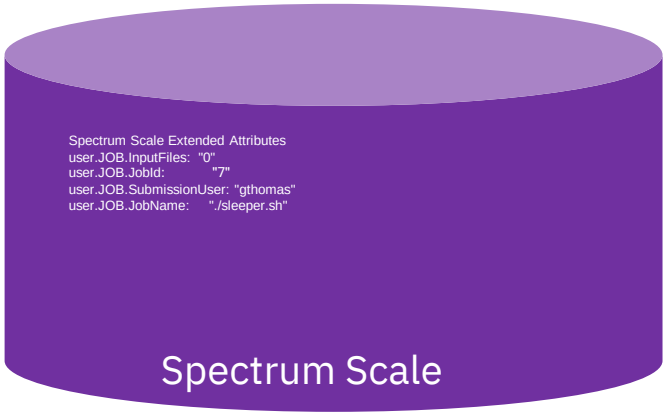
1. Run GATK4 pipeline

5. Leverage Provenance info

Data Scientist



2. Write Provenance data as XATTRS



3. Ingest system metadata

4. Extract and index provenance info

IBM

Spectrum LSF Provenance Data

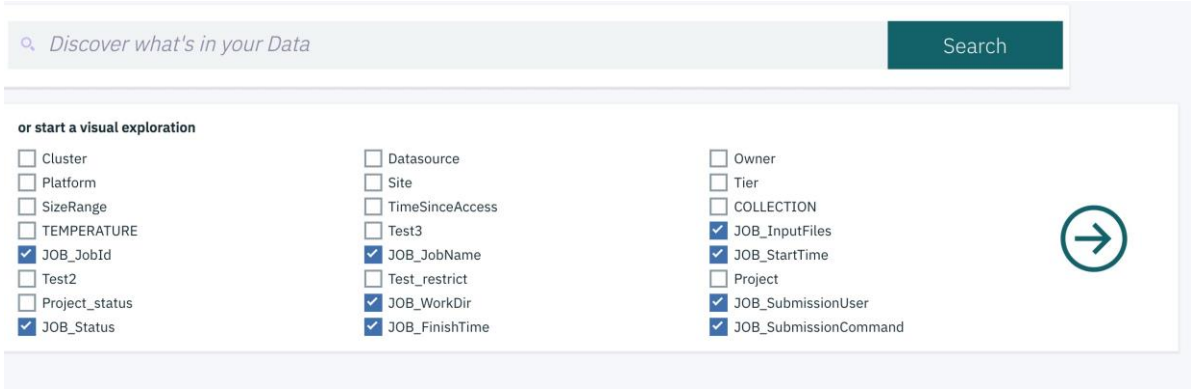
The extended attributes are the ones starting with "user.JOB"

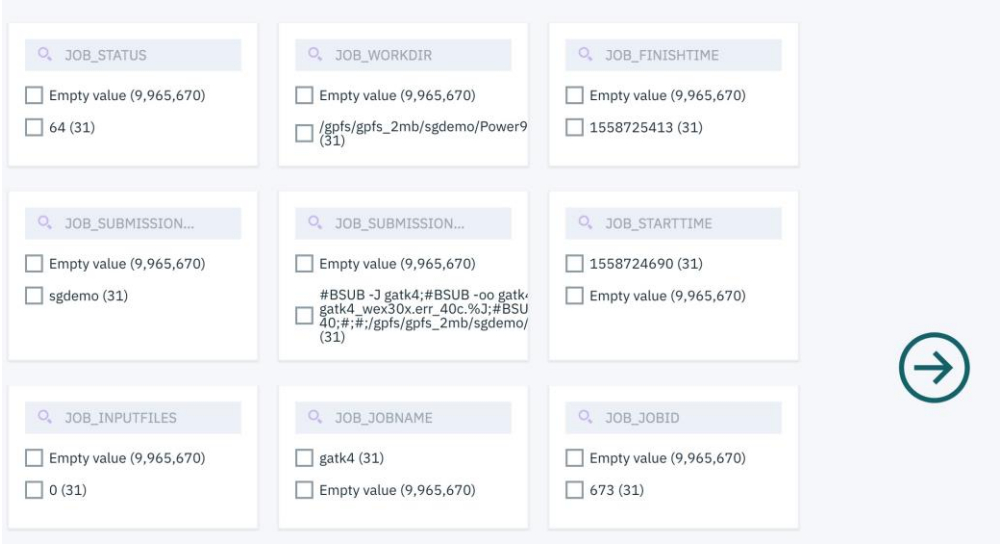
```
[gthomas@p95a07 ~]$ pwd
/home/gthomas
[gthomas@p95a07 ~]$ bsub -q gatk -o /gpfs/gpfs_2mb/gilbert/out2.txt -Ep
~/lsf/10.1/misc/examples/data_prov/tag.sh ./sleeper.sh
[gthomas@p95a07 ~]$ mmlsattr -d -L /gpfs/gpfs_2mb/gilbert/out2.txt
file name:            /gpfs/gpfs_2mb/gilbert/out2.txt
metadata replication: 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

IBM Spectrum Discover



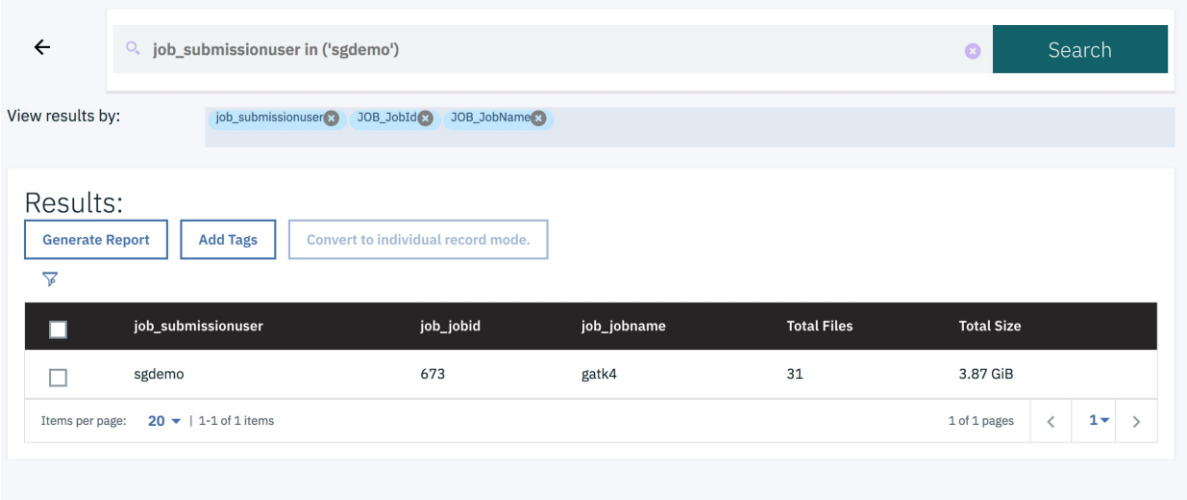


IBM

Spectrum LSF Provenance Data with Spectrum Discover

IBM Spectrum Discover

Grouped search results



IBM

Spectrum LSF Provenance Data with Spectrum Discover

IBM Spectrum Discover

View Individual Files Associated with workflow

Add Tags

filename	size	job_workdir	job_jobid	job_submissionuser	job_jobname	job_status	job_starttim
time_bwa.log	1038.000	/gpfs/gpfs_2mb/sgdemo/Power9/wes_30x	673	sgdemo	gatk4	64	155872469(
gcat_set_025_bwa.bam	1979566750.000	/gpfs/gpfs_2mb/sgdemo/Power9/wes_30x	673	sgdemo	gatk4	64	155872469(
time_Markduplicates.log	860.000	/gpfs/gpfs_2mb/sgdemo/Power9/wes_30x	673	sgdemo	gatk4	64	155872469(

IBM

Free 90-day

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

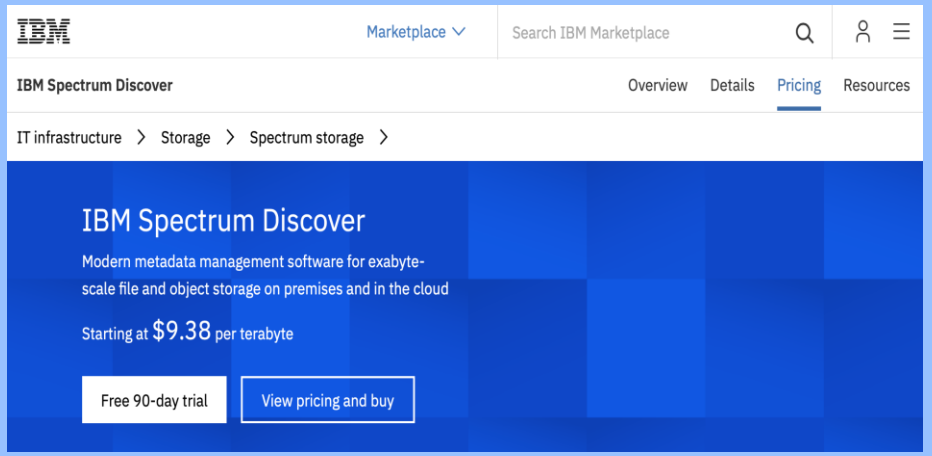
IBM Spectrum Discover Free Trial

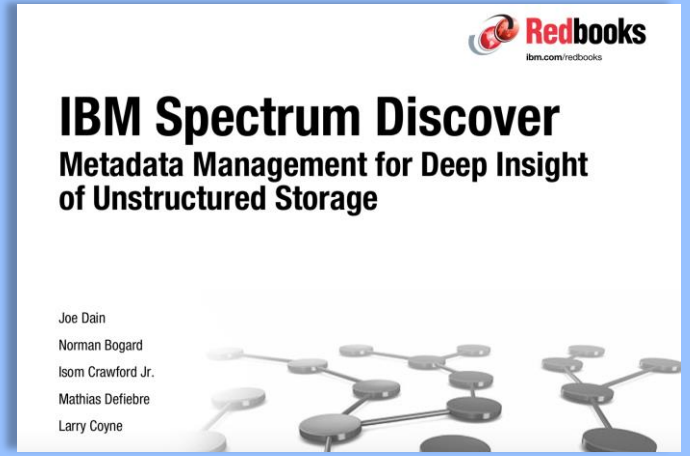
Unleash metadata-fueled insights for your unstructured data -- free for 90 days.

→ Free trial

www.ibm.com/marketplace/spectrum-discover

Learn more about Spectrum Discover





Web Page and Customer Resources

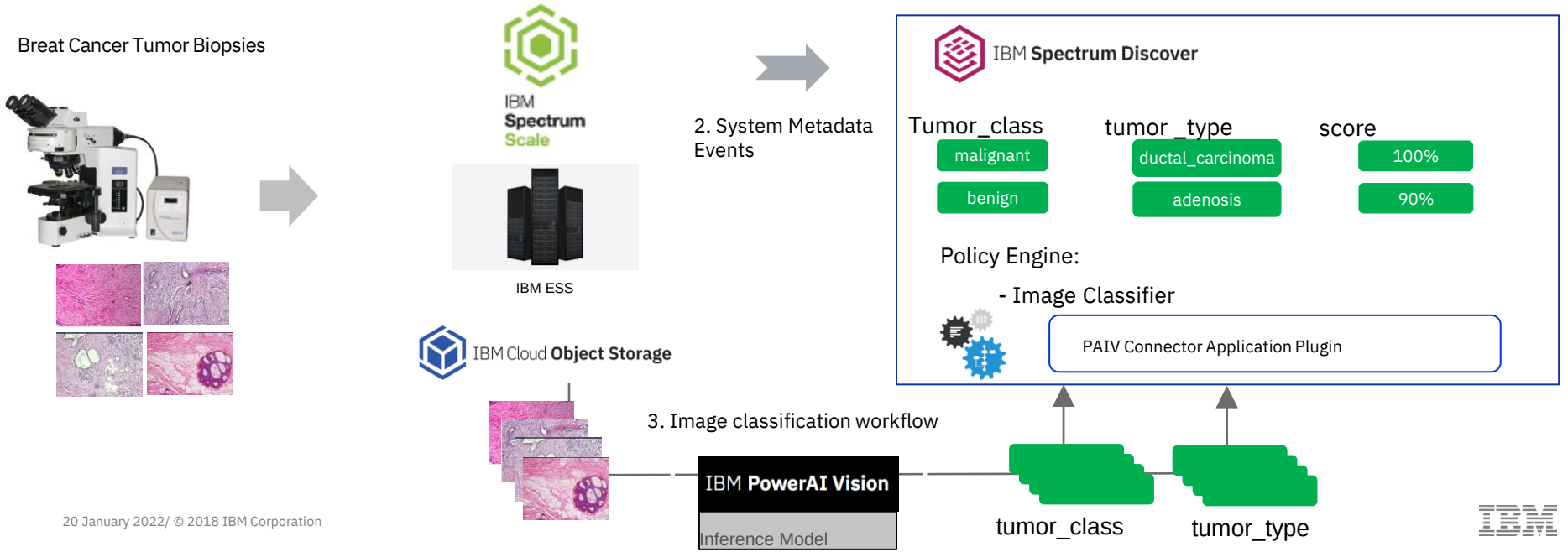
www.ibm.com/marketplace/spectrum-discover

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

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

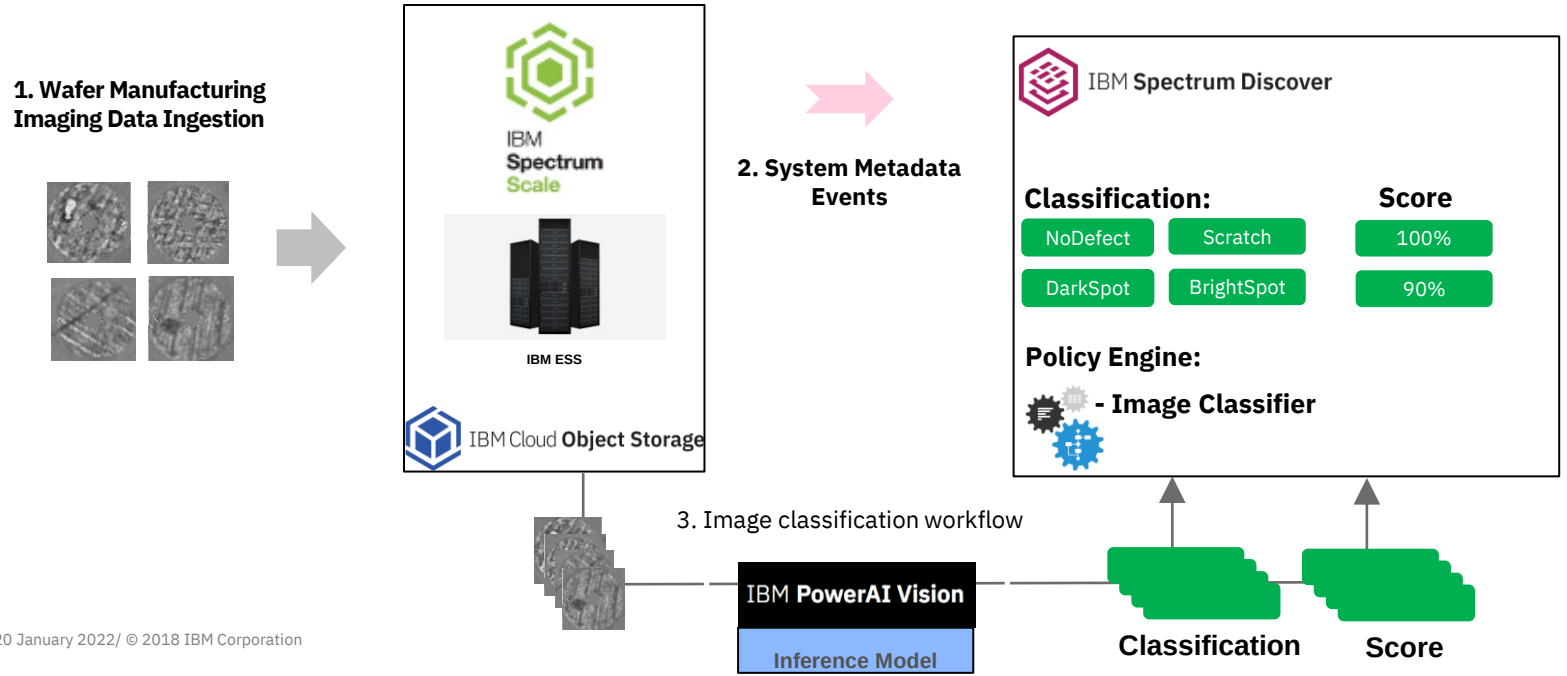
1. New imaging data ingested into Spectrum Scale / ESS, IBM COS storage
2. Storage sends Spectrum Discover system metadata events when new imaging data is ingested and Spectrum Discover builds catalog
3. Spectrum Discover policy automatically reads new imaging data from source storage, passes to the PowerAI Vision classification model, captures results and indexes into Spectrum Discover



Use Case: Automated Wafer Manufacturing Image Classification

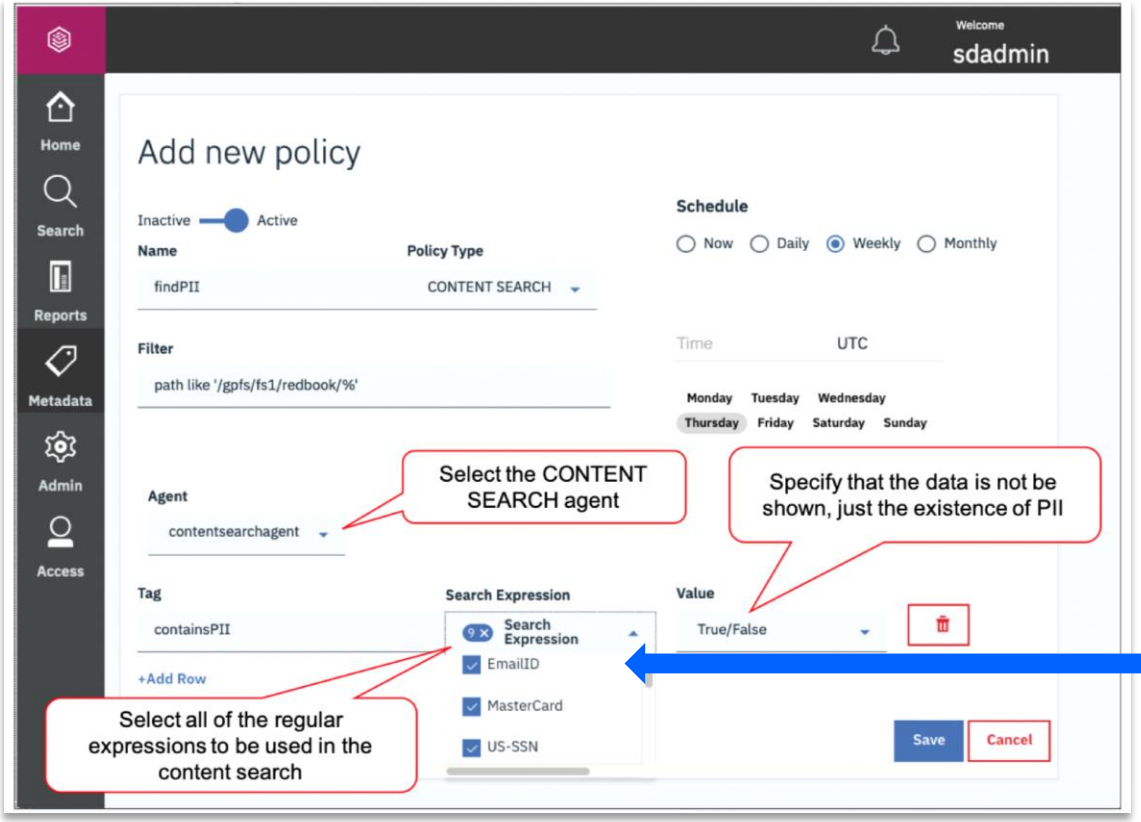
Event driven architecture to automatically classify and catalog wafer manufacturing data using PowerAI Vision inference model, Spectrum Discover, and Spectrum Scale / ESS / COS

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



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

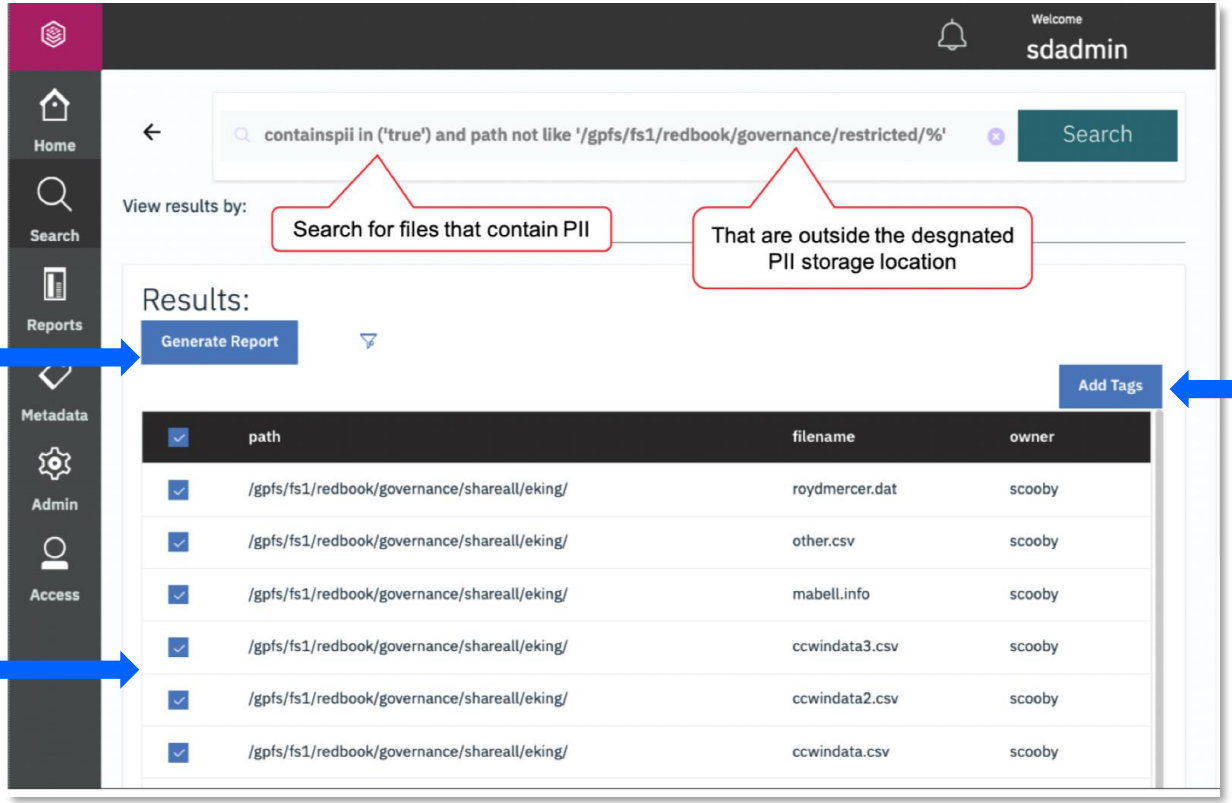
Data Mapping



Easily create your own custom search patterns

Discover your data with simple interface or report generation

Data Visualization



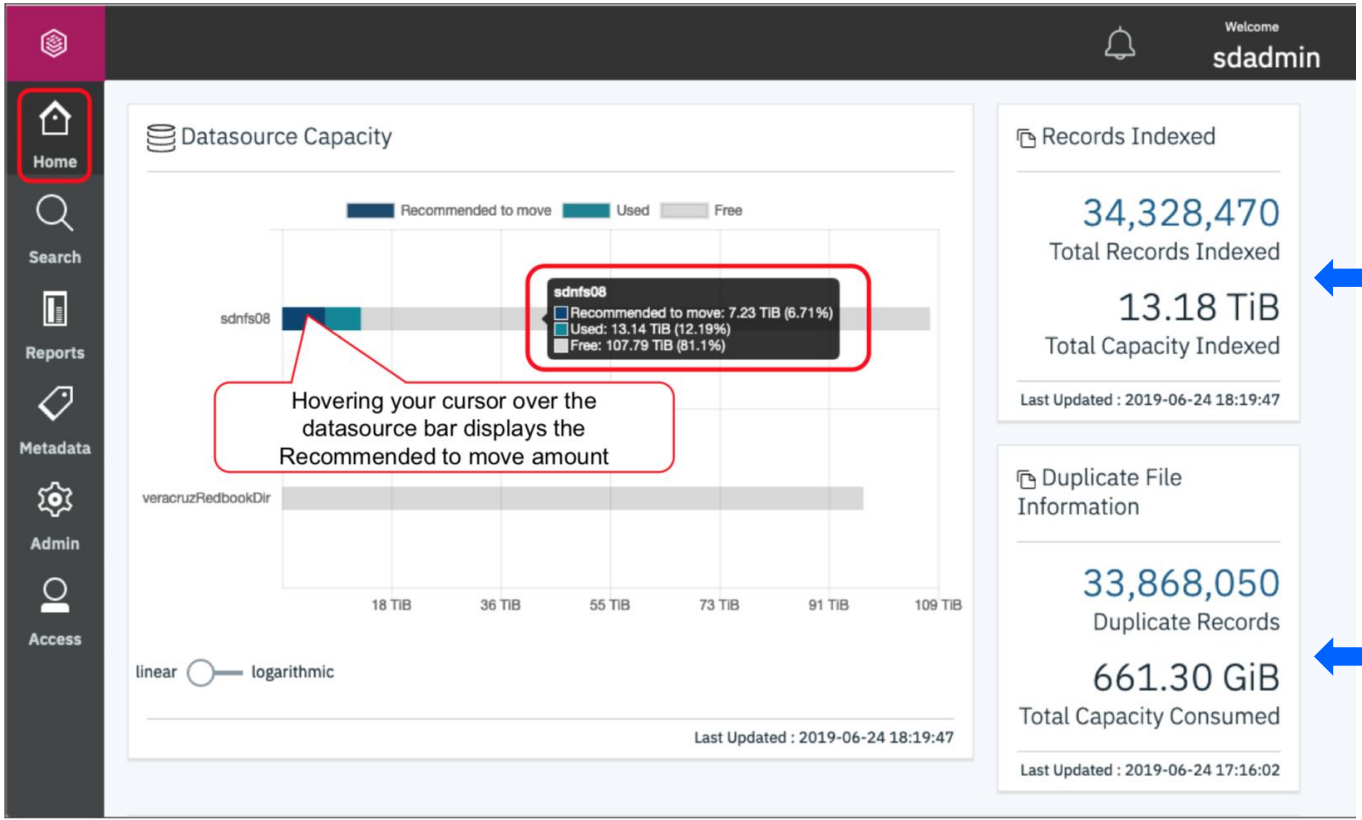
Generate reports

Drill down

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