IBM Spectrum Scale: Discover the value of your data with Spectrum Discover

—

Lars Lauber
Client Technical Specialist
Storage for Big Data and AI

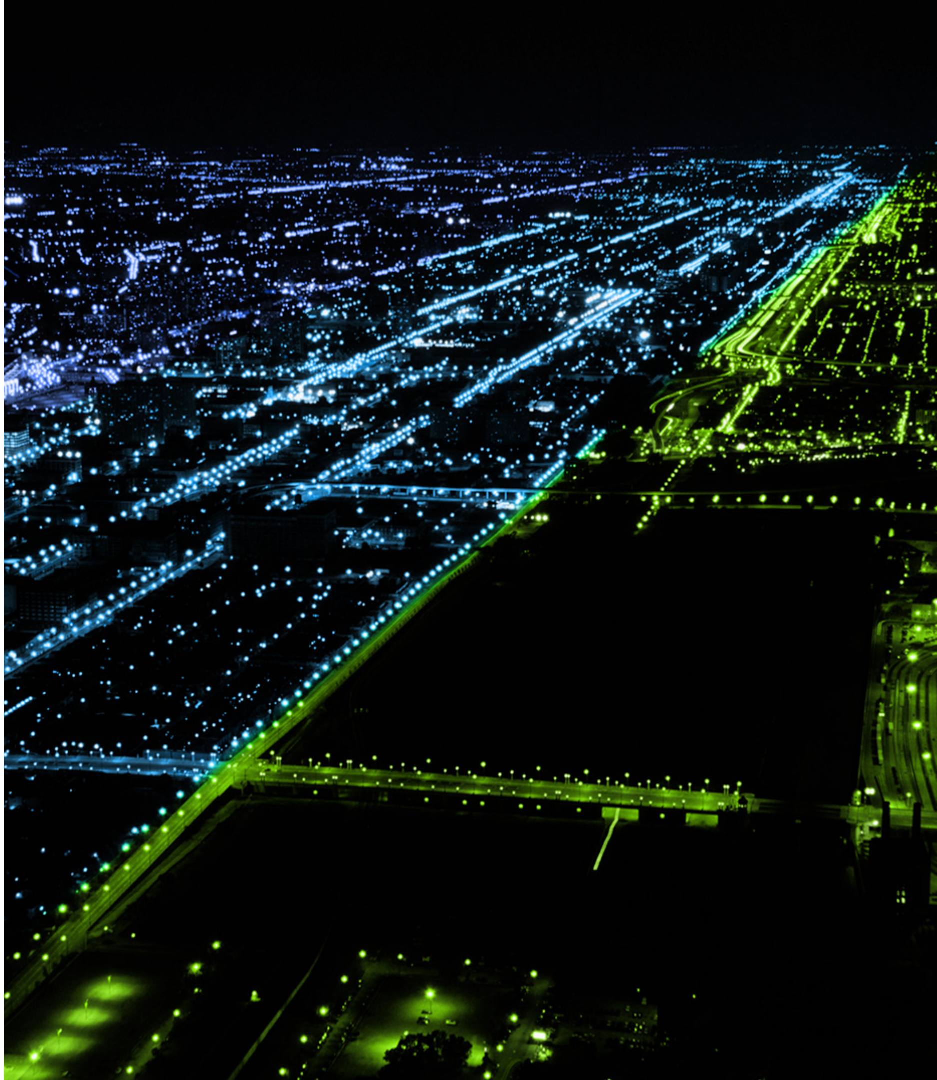
Harnessing the Value of Data

"The world's most valuable resource is no longer oil, but data."

The Economist, May, 2017

…how do companies ***harness the value?***

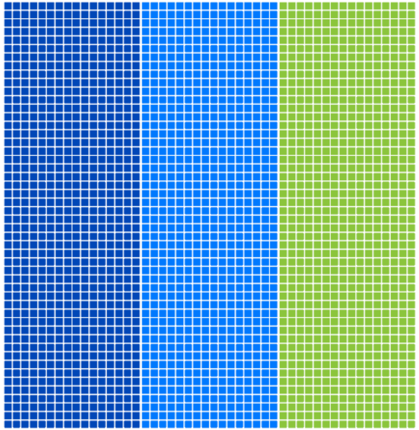
- Identify
- Categorize
- Utilize



Structured vs. Unstructured Data

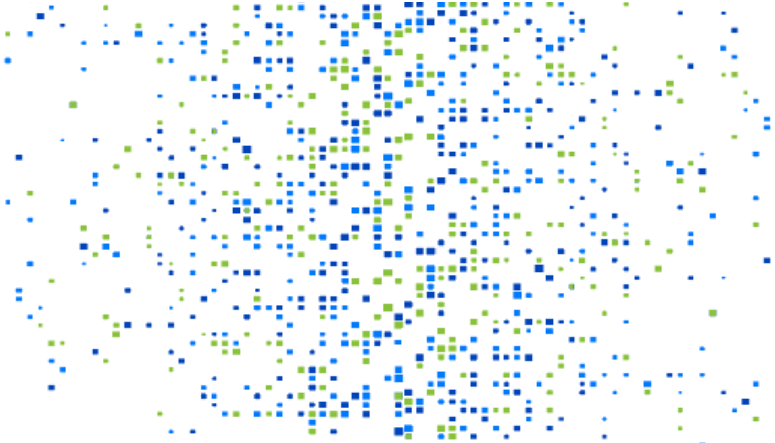
IBM Spectrum Discover

Structured



- Strictly organized, common schema
- Designed for management by computers
- Relational databases & spreadsheets
- Standard search operations

Unstructured



- No uniform structure
- Designed for use by humans & devices
- Word docs, PDFs, emails, videos, IoT sensor data, audio files, emails, HTML, & images
- Limited data visibility

Unstructured Data Is Hard to Manage

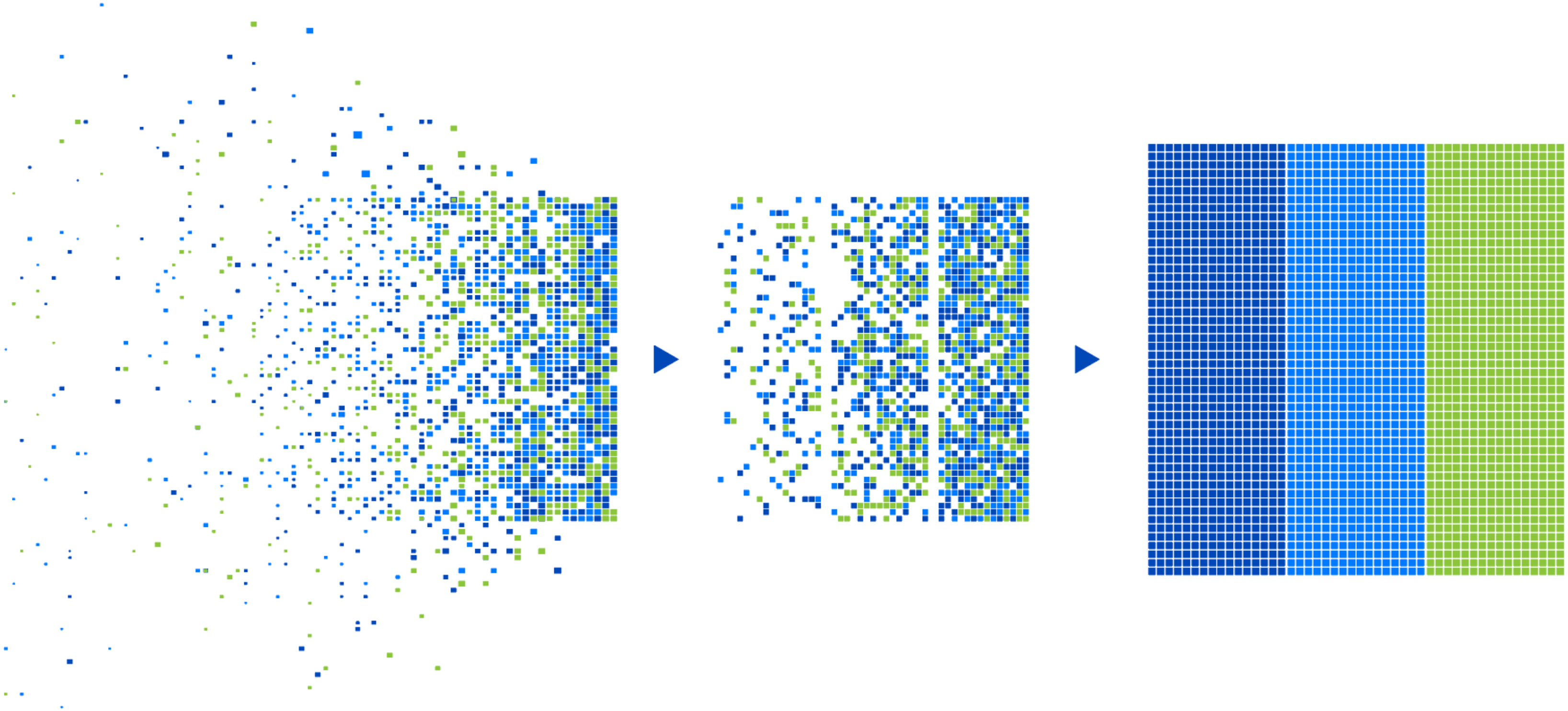
For exabyte-scale data stores…

- 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

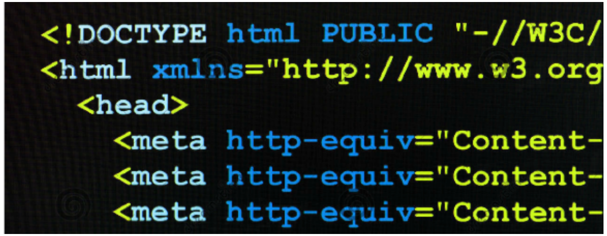
What is needed?

IBM Spectrum Discover

The ability to bring structure to unstructured data.



Metadata is the Key to Data Organization & Insight



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

Three Types of Metadata

1

Metadata can come from a system

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

2

Metadata can be custom

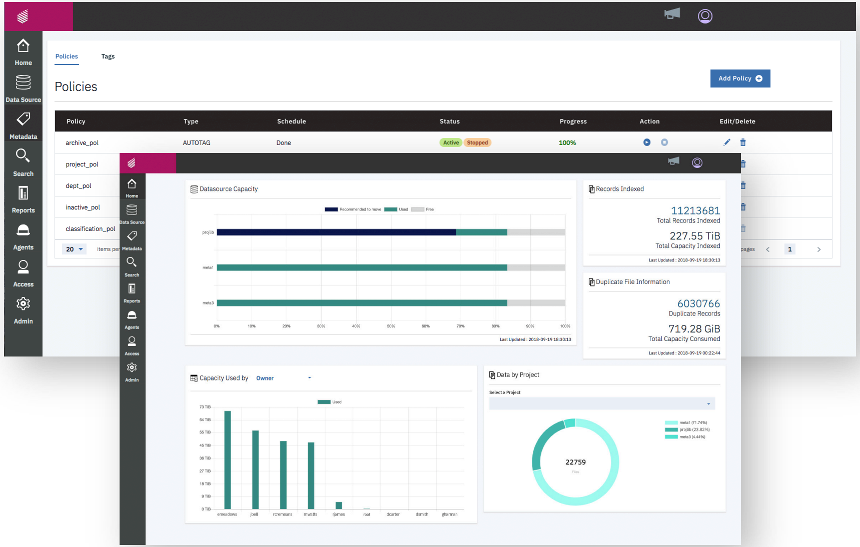
(map to various business & scientific aspects)

3

Metadata can be derived from analytics

(percent confident)

IBM **Spectrum Discover**



Policies

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

20 items per … pages 1

Datasource Capacity

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

Records Indexed

11213681
Total Records Indexed

227.55 TiB
Total Capacity Indexed

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

Duplicate File Information

6030766
Duplicate Records

719.28 GiB
Total Capacity Consumed

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

Capacity Used by Owner

Data by Project

Select a Project

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

IBM Spectrum Discover Accelerates Customer Value

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

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

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

Spectrum Discover use cases

	Use Cases	Curate	Identify	Analyze/Comply/Optimize
 Data Scientist	Analytics	Curate Metadata	Identify relevant datasets	Tag results
 Data Steward	Risk Mitigation		Identifiy sensitive data	Classify files and objects
 Storage Admin	Optimization - archival/tiering		Identify aged data	Tag archival data
 Storage Admin	Optimization - data removal		Identify/verify/tag duplicate data	Tag/ Delete duplicate data

IBM Spectrum Discover overview

IBM Spectrum Discover

Where

Backup, File & Object Storage

IBM Spectrum Archive

IBM Spectrum Scale

IBM Cloud Object Storage

IBM Spectrum Protect

IBM Elastic Storage Server

ISILON

amazon web services™ | S3

NetApp

ceph

What

Index & Tag Big Data

IBM Spectrum Discover

Search

Reporting

Dashboard

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

Why

High Speed Data Insight

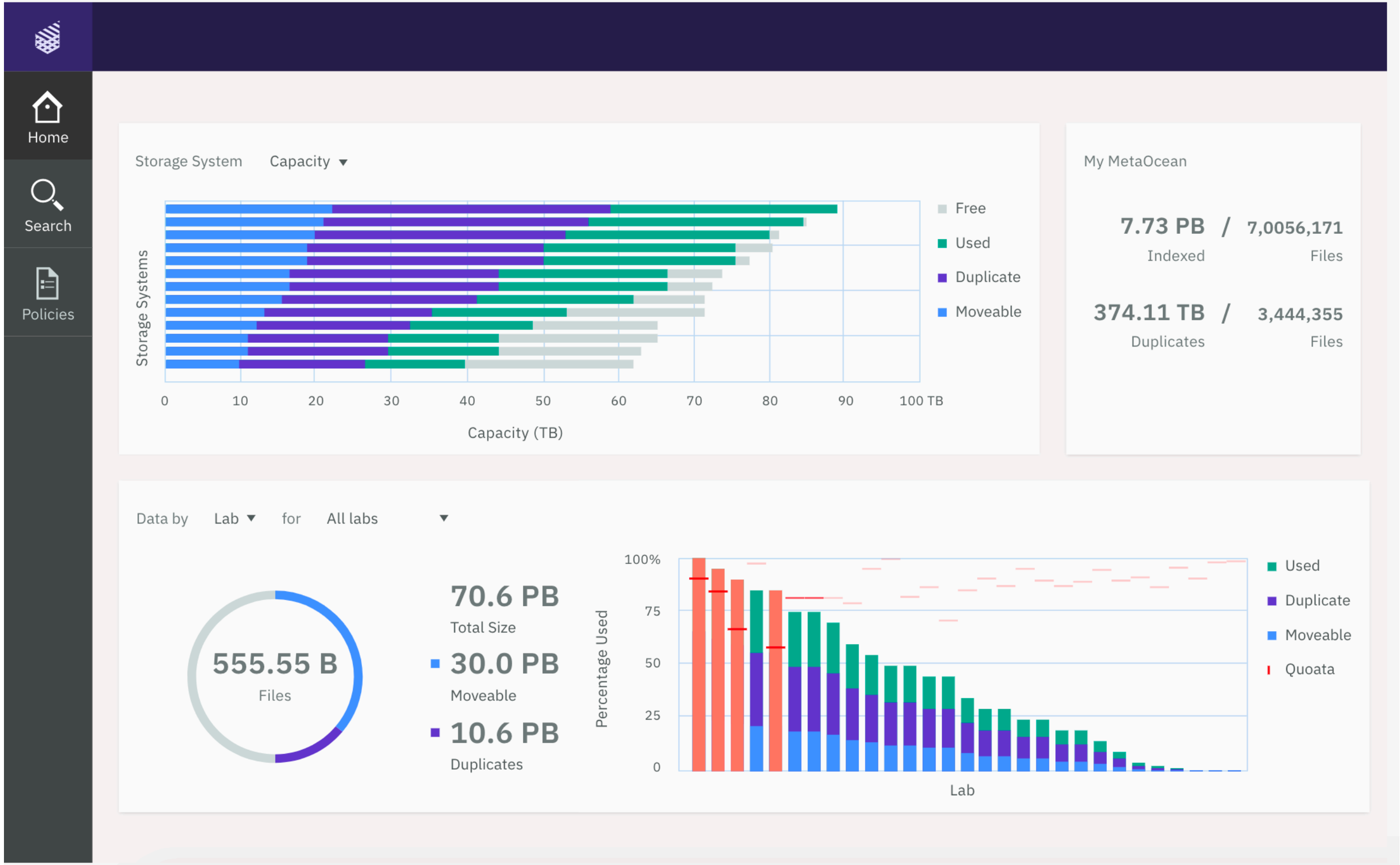
Large-Scale Analytics and AI/ML
- Data discovery
- Dataset identification
- Data pipeline progression

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

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

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

Capacitymanagement and duplicates:



My MetaOcean

7.73 PB / 7,0056,171
Indexed Files

374.11 TB / 3,444,355
Duplicates Files

555.55 B
Files

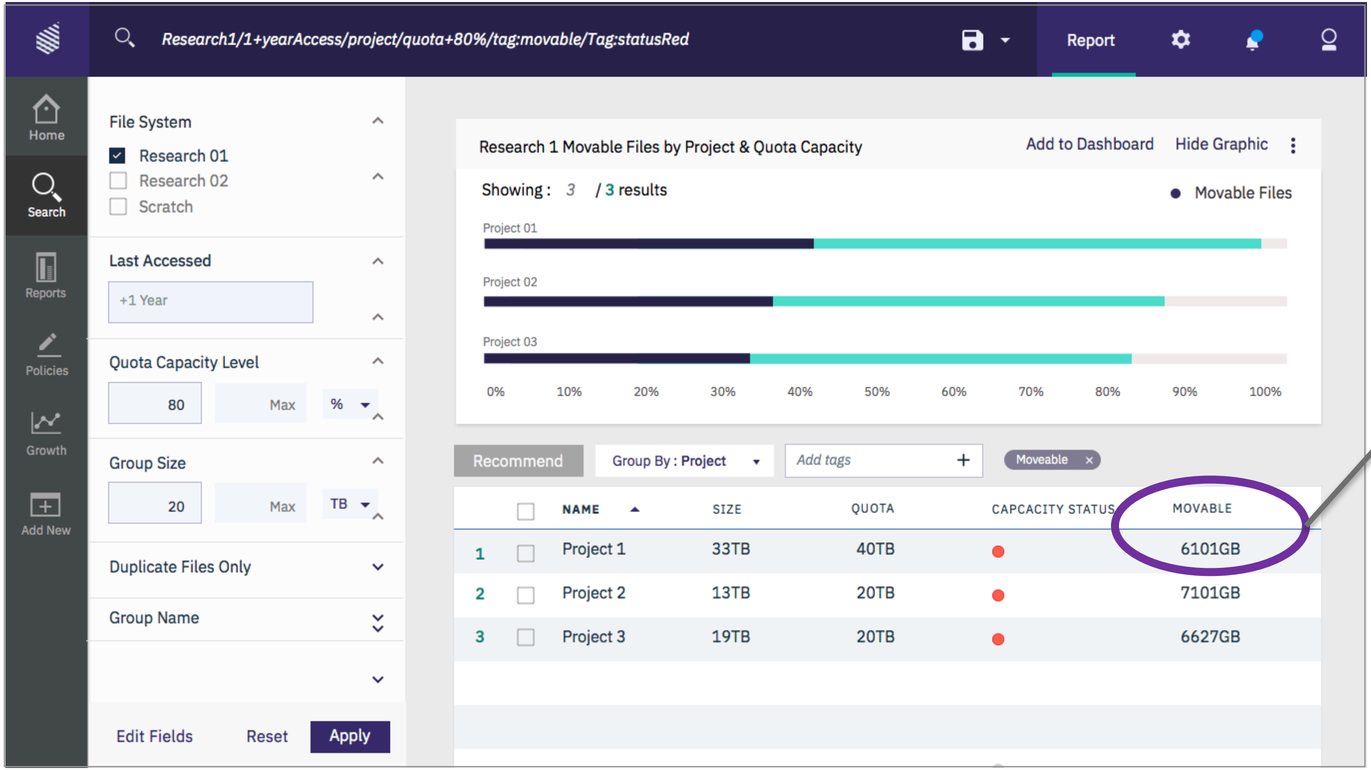
70.6 PB
Total Size

30.0 PB
Moveable

10.6 PB
Duplicates

Prepare to move data

IBM Spectrum Discover



Research1/1+yearAccess/project/quota+80%/tag:movable/Tag:statusRed

File System
- ☑ Research 01
- □ Research 02
- □ Scratch

Last Accessed: +1 Year

Quota Capacity Level: 80 – Max %

Group Size: 20 – Max TB

Duplicate Files Only

Group Name

Research 1 Movable Files by Project & Quota Capacity

Showing : 3 / 3 results

Project 01

Project 02

Project 03

Group By : Project

Add tags +

Moveable ×

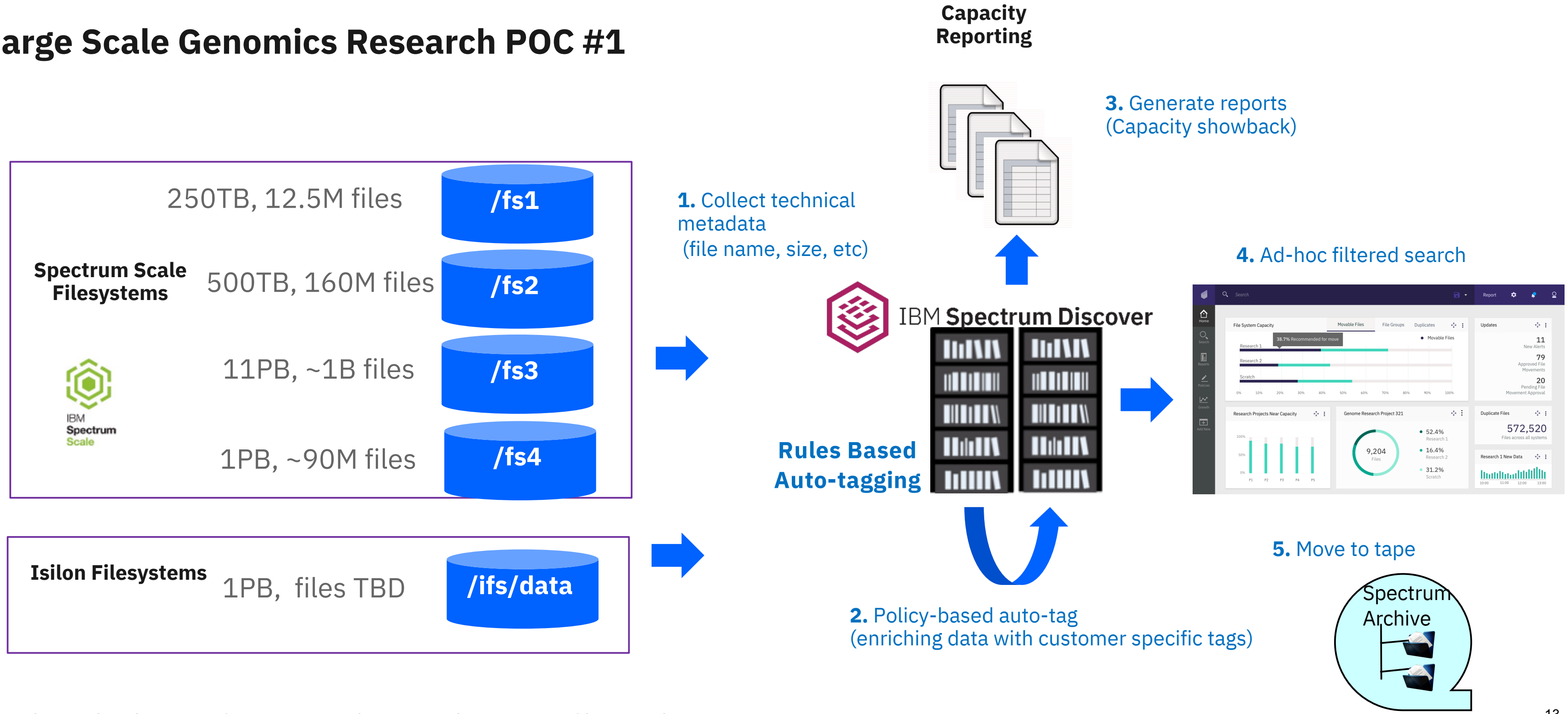
		NAME	SIZE	QUOTA	CAPCACITY STATUS	MOVABLE
1	□	Project 1	33TB	40TB	🔴	6101GB
2	□	Project 2	13TB	20TB	🔴	7101GB
3	□	Project 3	19TB	20TB	🔴	6627GB

prepares recommendations to move data

IBM Spectrum Discover

Use Case: Tiering of File/Objects

Large Scale Genomics Research POC #1



Capacity Reporting

3. Generate reports (Capacity showback)

Spectrum Scale Filesystems

IBM Spectrum Scale

250TB, 12.5M files — /fs1

500TB, 160M files — /fs2

11PB, ~1B files — /fs3

1PB, ~90M files — /fs4

Isilon Filesystems

1PB, files TBD — /ifs/data

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

IBM Spectrum Discover

Rules Based Auto-tagging

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

4. Ad-hoc filtered search

5. Move to tape

Spectrum Archive

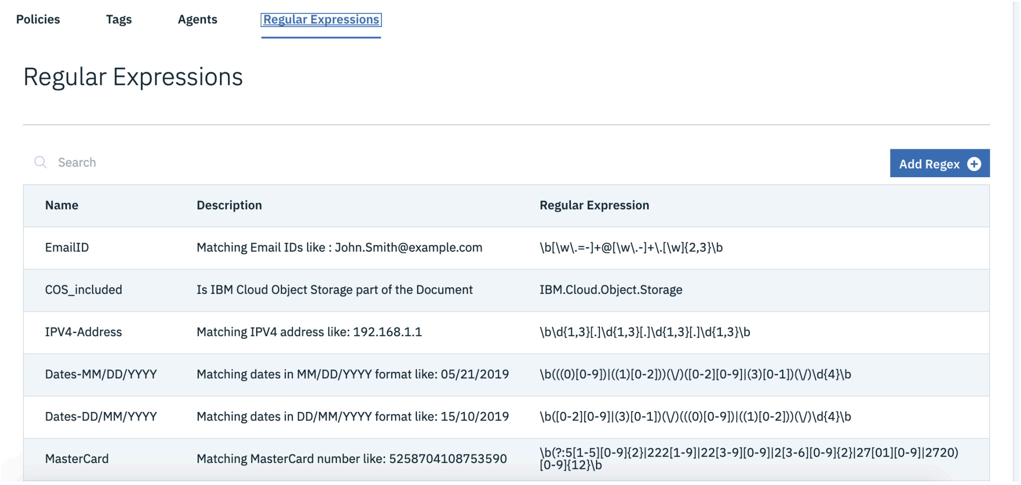
Special Action Agent: Contentsearch

Contentsearch is based on Apache Tika, which can detect and extract metadata out of different file types

In this instance, the Tika service acts as the Action Agent, although it‘s already built into Spectrum Discover (>2.0.1)

Analysis is done based on Regex

Can be used to detect Personal Identifiable Information (PII), such as Credit Cards or E-Mail adresses

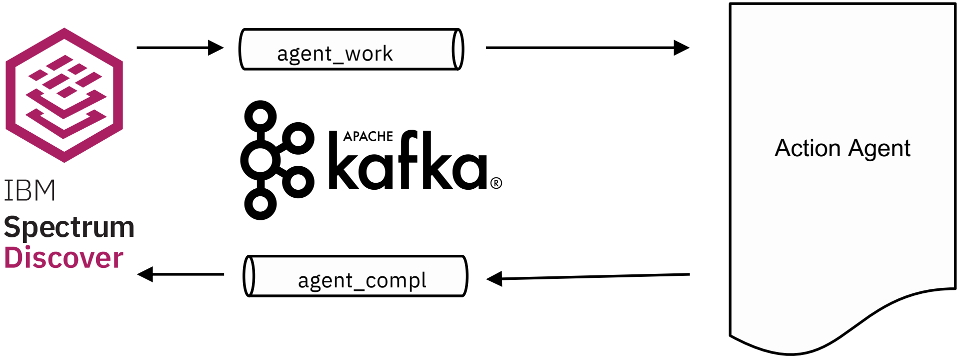


Regular Expressions

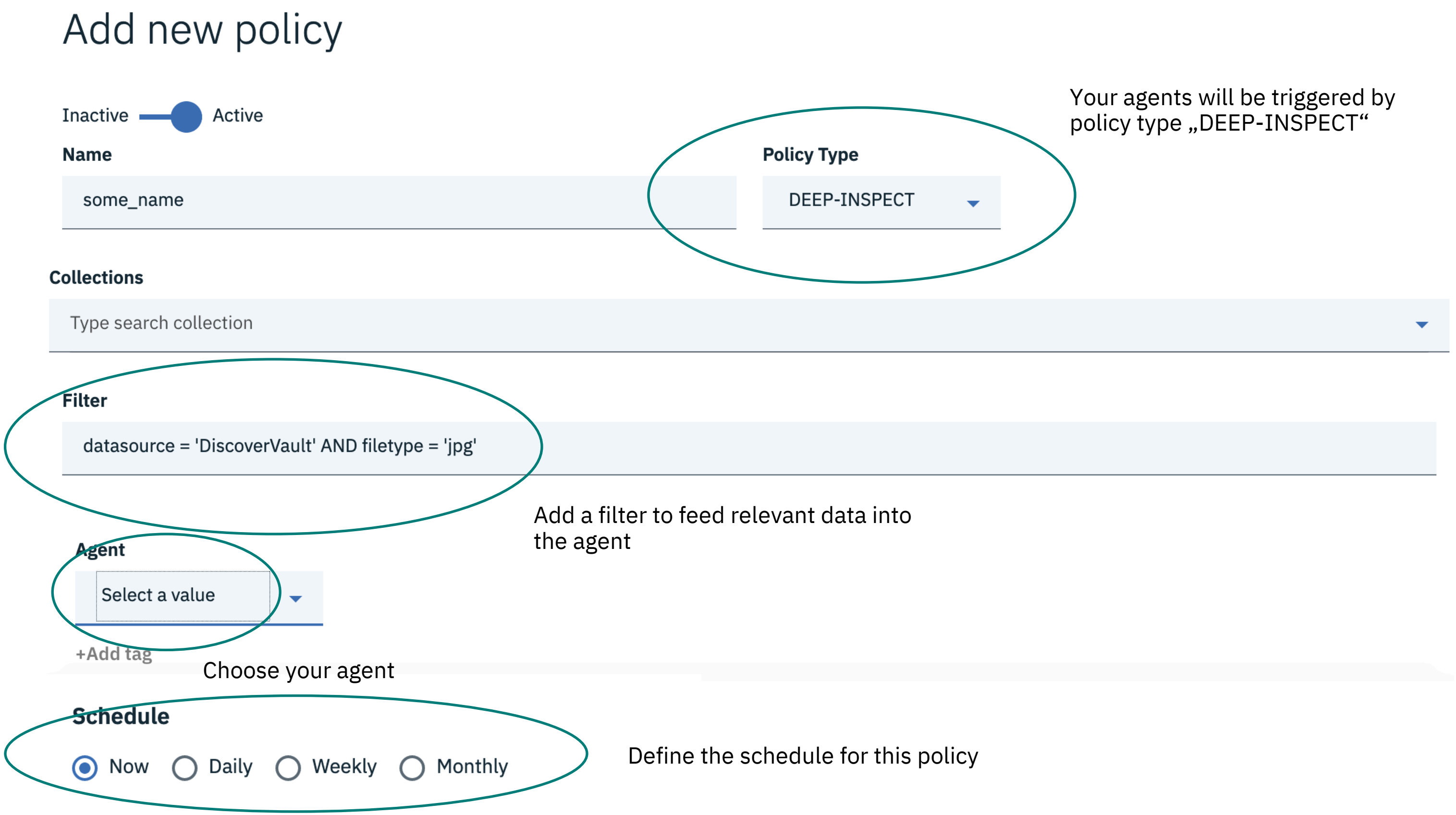
Name	Description	Regular Expression
EmailID	Matching Email IDs like : John.Smith@example.com	\b[\w\.=-]+@[\w\.-]+\.[\w]{2,3}\b
COS_included	Is IBM Cloud Object Storage part of the Document	IBM.Cloud.Object.Storage
IPV4-Address	Matching IPV4 address like: 192.168.1.1	\b\d{1,3}[.]\d{1,3}[.]\d{1,3}[.]\d{1,3}\b
Dates-MM/DD/YYYY	Matching dates in MM/DD/YYYY format like: 05/21/2019	\b(((0)[0-9])|((1)[0-2]))(\/)([0-2][0-9]|(3)[0-1])(\/)\d{4}\b
Dates-DD/MM/YYYY	Matching dates in DD/MM/YYYY format like: 15/10/2019	\b([0-2][0-9]|(3)[0-1])(\/)(((0)[0-9])|((1)[0-2]))(\/)\d{4}\b
MasterCard	Matching MasterCard number like: 5258704108753590	\b(?:5[1-5][0-9]{2}|222[1-9]|22[3-9][0-9]|2[3-6][0-9]{2}|27[01][0-9]|2720)[0-9]{12}\b

Action Agent workflow

1. File filtering by Policy in Discover
2. Discover sends requests in Kafka pipeline
3. Action Agent does some magic
4. Action Agent responds to Kafka pipeline
5. Discover updates metadata on files



Define policy



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

Your agents will be triggered by policy type „DEEP-INSPECT“

Add a filter to feed relevant data into the agent

Choose your agent

Define the schedule for this policy

What happens next?

- Discover sends matching data entries into work pipeline
- Action Agent watches work pipeline for new entries
- Action Agent extracts datapaths from the message
- Action Agent does some magic
- Action Agent builds response and sends it to compl pipeline
- Spectrum Discover adds metadata to the database

Sample message work pipeline

```
{
  "mo_ver": "1.0",
  "action_id": "DEEPINSPECT",
  "action_params": {
    "agent": "extractagent",
                "tags": {"extract_tags": ["vin", "sensor"]}
  },
  "agent": " extractagent",
  "policy_id": " extractpol",
  "docs": [
    {"path": "/fs1/path1/file1.txt", "fkey": "spectrumscale.cluster.example"},
    {"path": "/fs1/path1/file2.txt", "fkey": "spectrumscale.cluster.example"},
                ......
                {"path": "/fs1/path1/file3.txt", "fkey": "spectrumscale.cluster.example"}
   ]
}
```

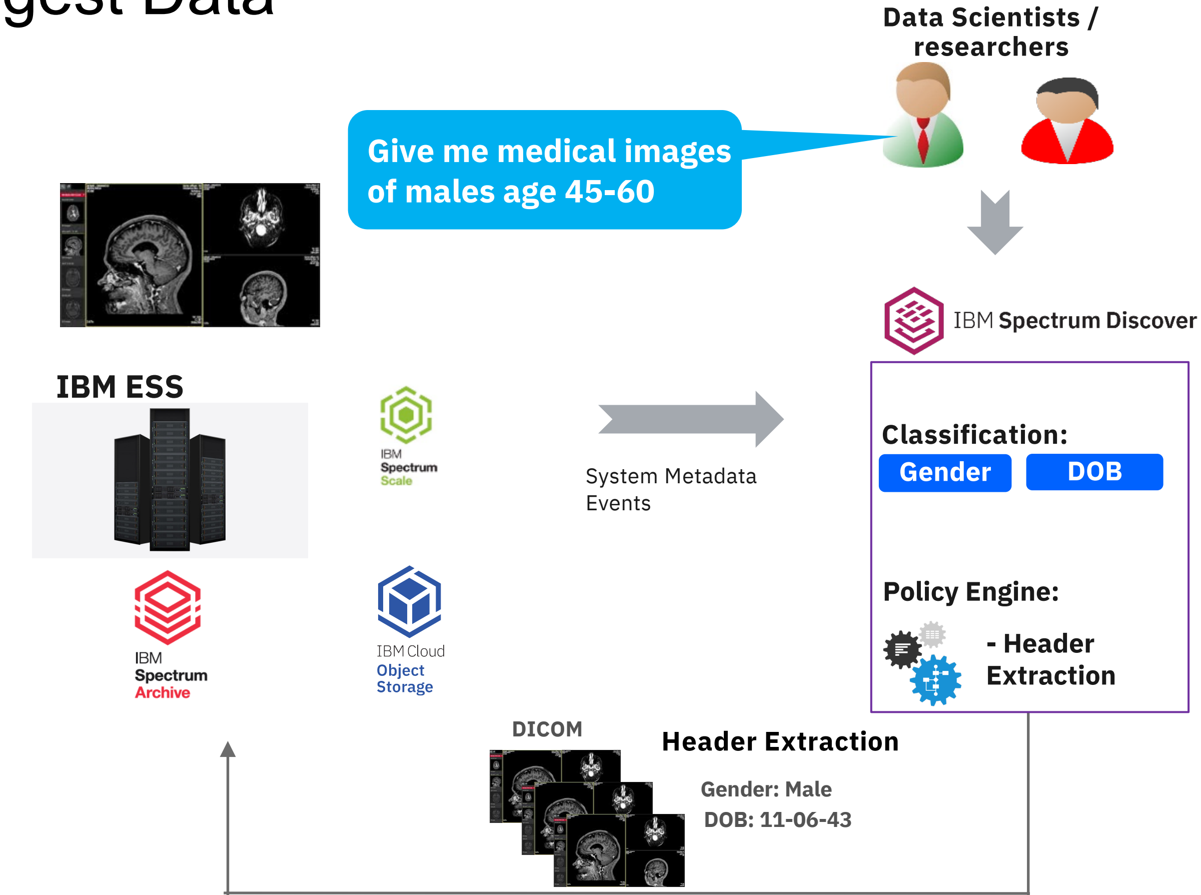
Sample message compl pipeline

```
{
  "mo_ver": "1.0",
  "policy_id": " extractpol",
  "docs": [
     {"status": "success", "tags": {"vin": "vin-value", "sensor": "sensor-value"}, "path":
"/fs1/path1/file1.txt", "fkey": "spectrumscale.cluster.example "},
     {"status": "success", "tags": {"vin": "vin-value", "sensor": "sensor-value"}, "path":
"/fs1/path1/file1.txt", "fkey": "spectrumscale.cluster.example "},
     {"status": "failed", "tags": {}, "path": "/fs1/path1/file1.txt", "fkey":
"spectrumscale.cluster.example "}

   ]
}
```

Use Case: Tag, Search, Ingest Data

IBM Spectrum Discover



Using Spectrum Discover, Data Scientists can:

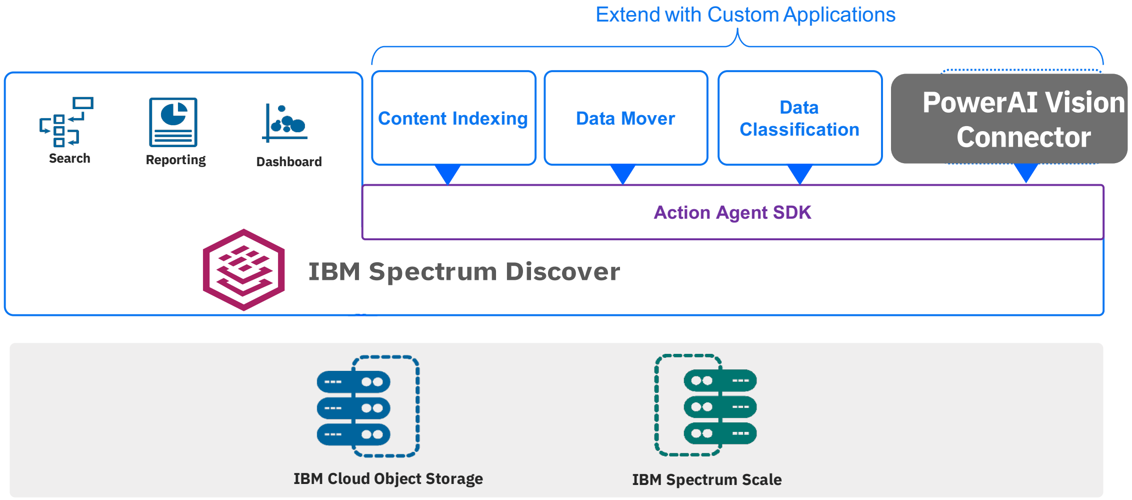
Add custom tags to files to match project, department, data owner, etc.

Use deep inspection API's to index specific content in the Spectrum Discover database.

Find similar projects

Spectrum Discover – App Catalogue

IBM Spectrum Discover



Spectrum Discover Application

Reads images from Spectrum Scale and / or COS, does some specific task to capture metadata, and updates Spectrum Discover catalog with results

GitHub

docker hub

IBM / Spectrum_Discover_App_Catalog

IBM Storage for AI: Data analysis, discovery, ML and AI workflows

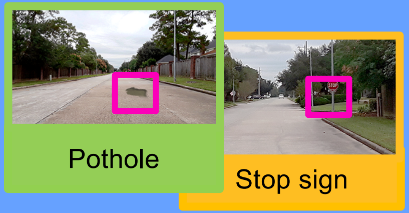
IBM Spectrum Discover

Ingest

IBM Cloud Object Storage

IBM Spectrum Scale

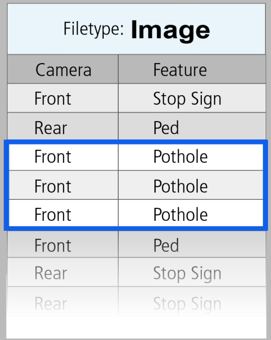
Data mapping, data visualization, data discovery



IBM PowerAI Vision

IBM Spectrum Discover

Curate, organize & data activation



Filetype: **Image**

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

IBM Spectrum Discover

Train

Apache Spark

TensorFlow

IBM PowerAI Vision

IBM Cloud Object Storage

IBM Spectrum Scale

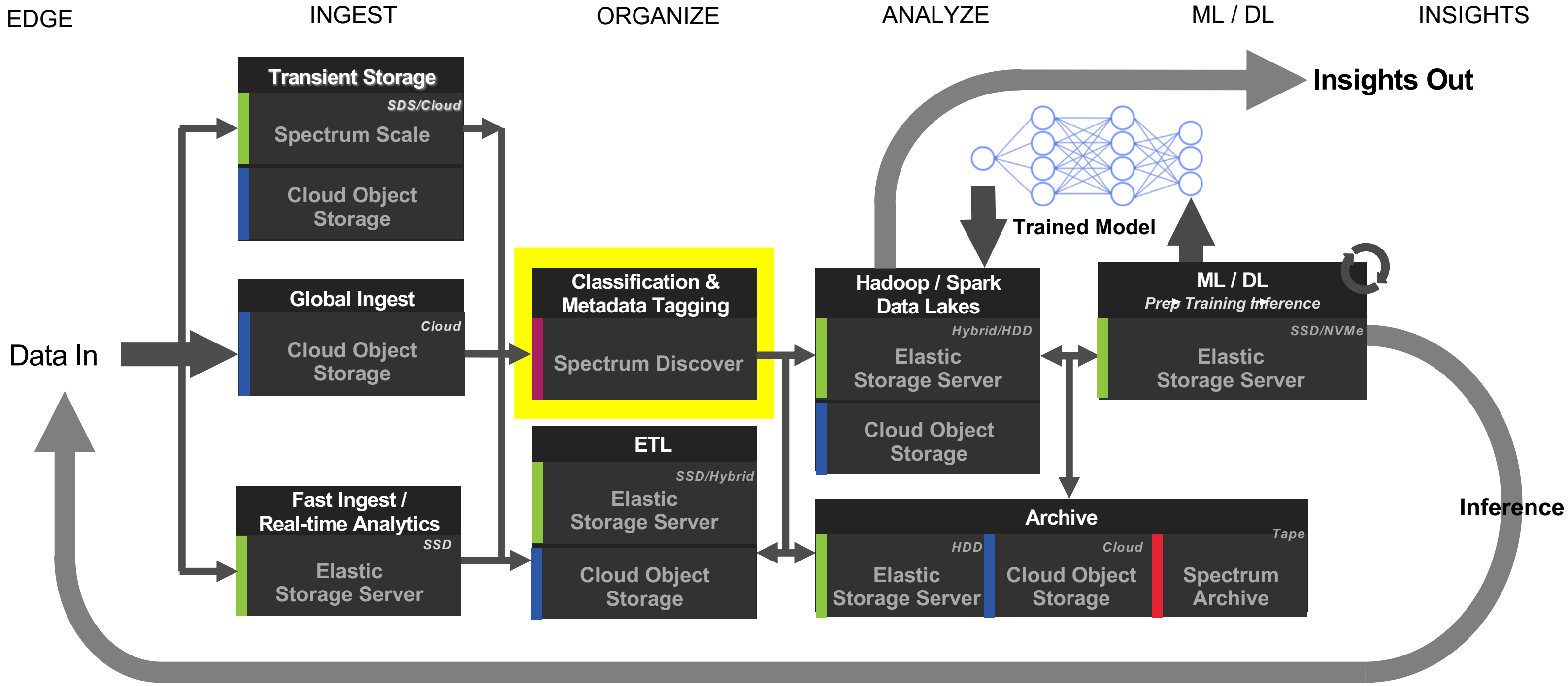
User / data admin

IT admin / architect

Data scientist

IBM Storage improves data science productivity across the entire data pipeline

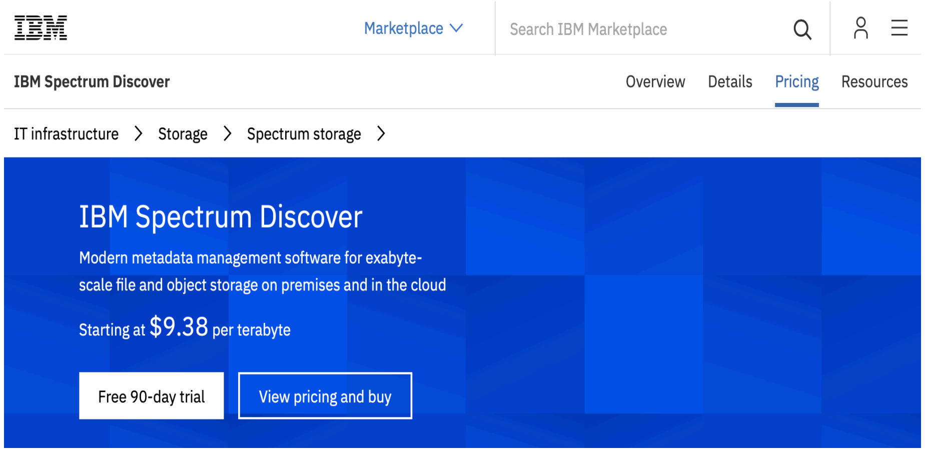
IBM Spectrum Discover



IBM Spectrum Scale IBM Cloud Object Storage IBM Spectrum Discover IBM Spectrum Archive

Learn more about Spectrum Discover

IBM Spectrum Discover



IBM

Marketplace

Search IBM Marketplace

IBM Spectrum Discover

Overview Details Pricing Resources

IT infrastructure > Storage > Spectrum storage >

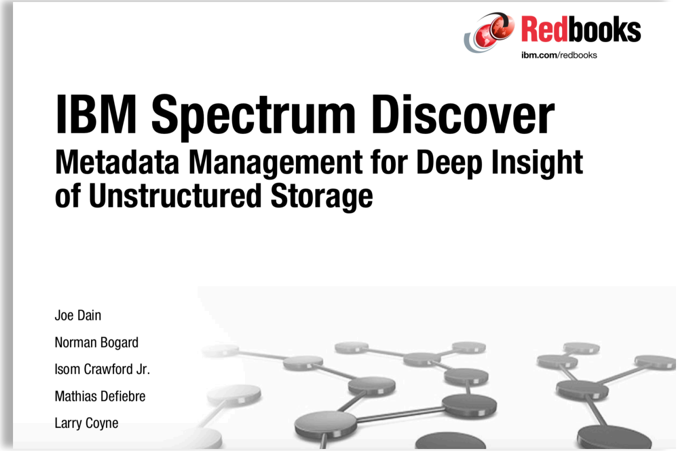
IBM Spectrum Discover

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



Redbooks
ibm.com/redbooks

IBM Spectrum Discover

Metadata Management for Deep Insight of Unstructured Storage

Joe Dain
Norman Bogard
Isom Crawford Jr.
Mathias Defiebre
Larry Coyne

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

Web Page and Customer Resources

www.ibm.com/marketplace/spectrum-discover