IBM **Spectrum Discover**

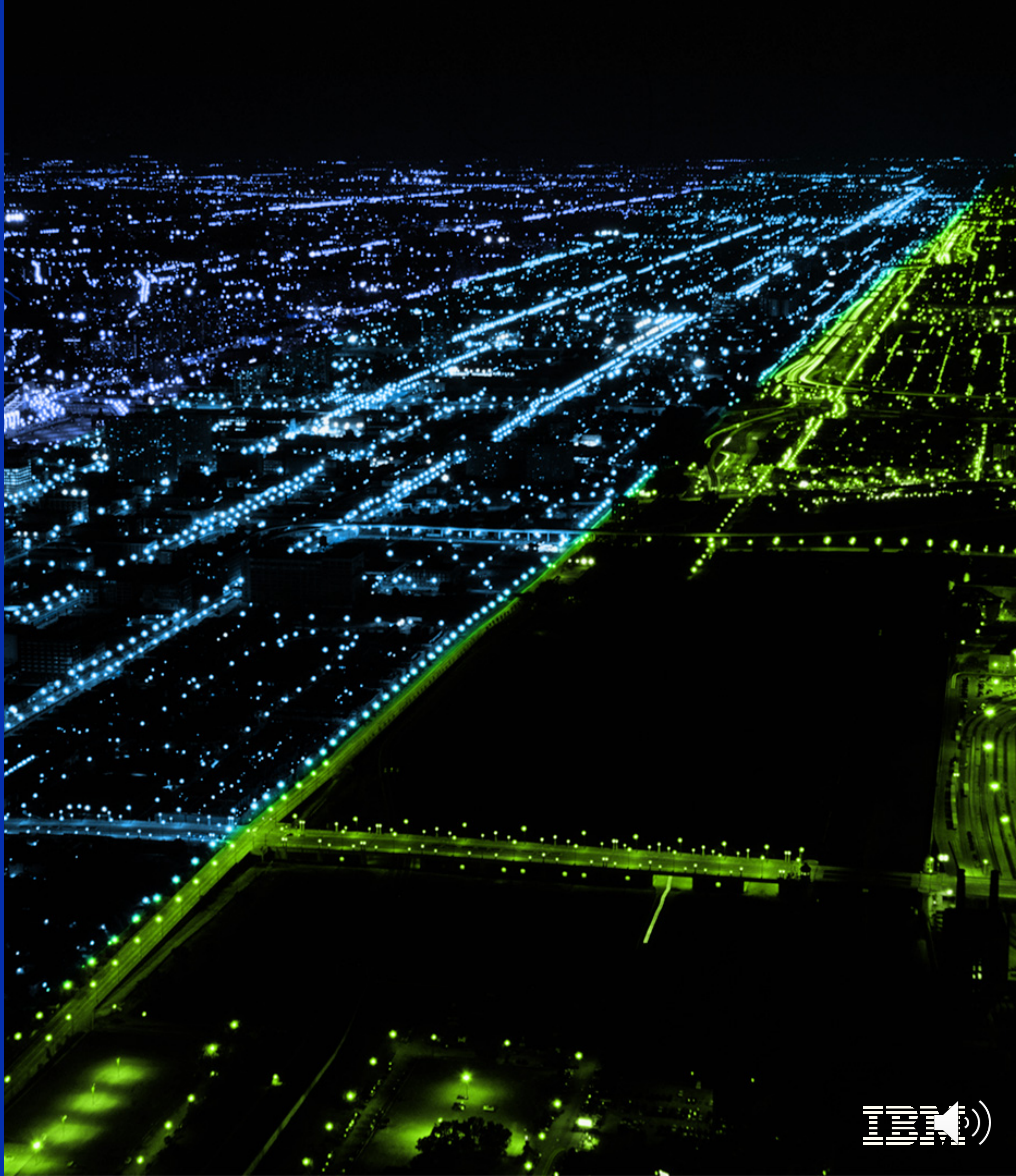
Exceptional Data Insight for Petabyte-Scale Unstructured Data Storage

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**



Number of enterprises with 1,000 TB+ unstructured data stores grew

3X

from 2016 to 2017

39%

of firms see sourcing, gathering, managing & governing data as their biggest challenges when using systems of insight

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

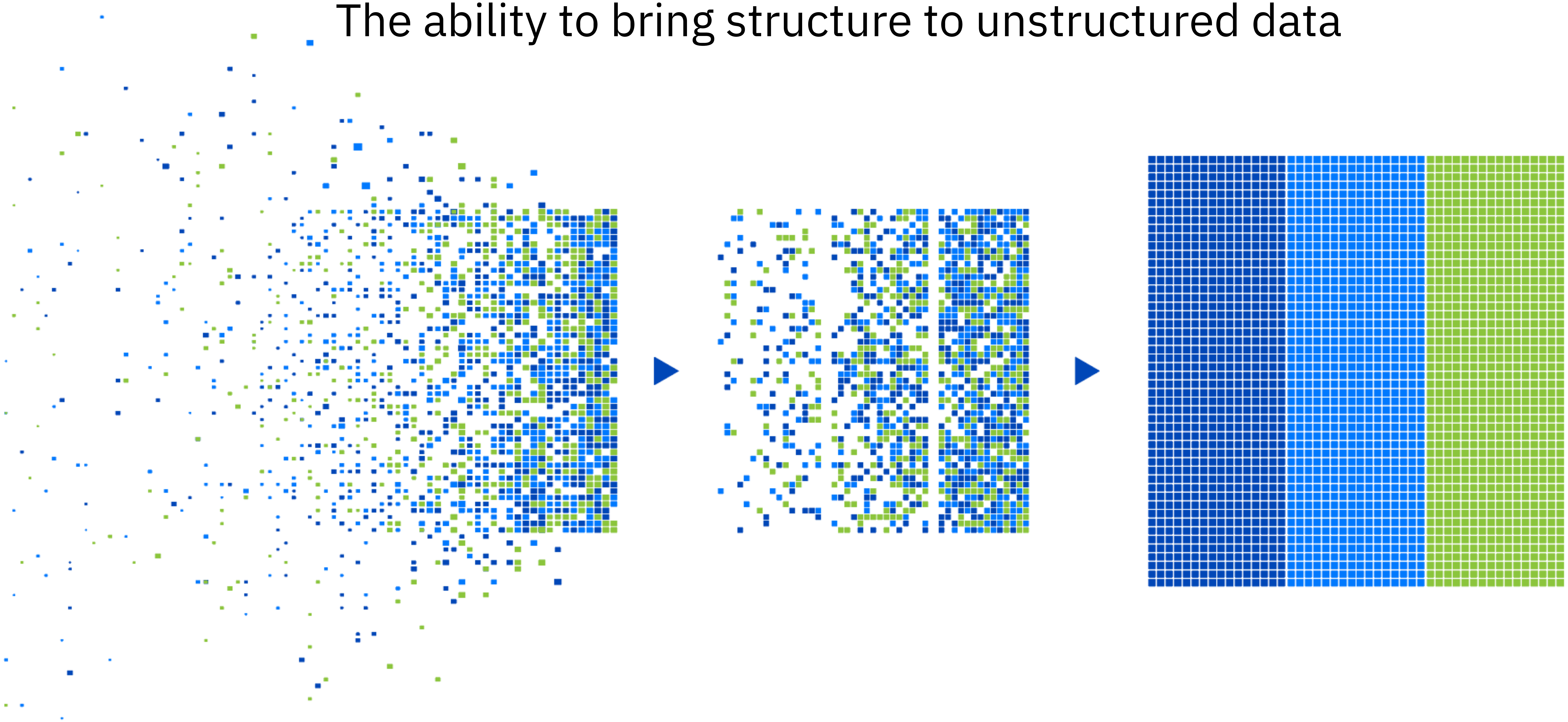
Unstructured Data is Hard to Manage

For 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?

The ability to bring structure to unstructured data



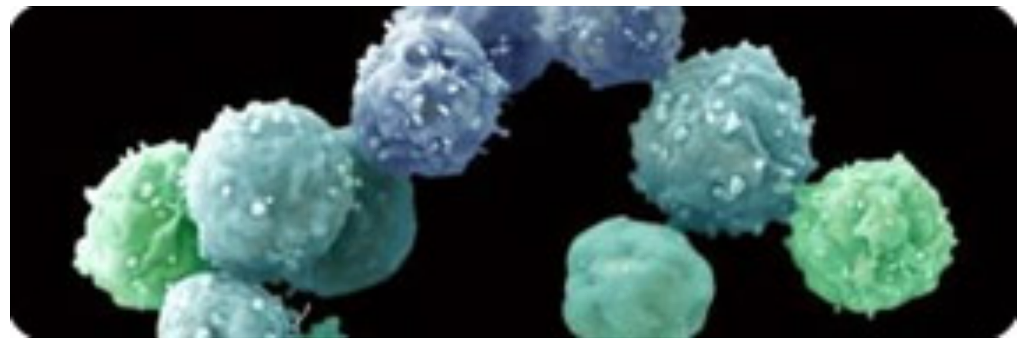
Metadata: Key to Unlocking Data Value & Improving Management

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

System Metadata

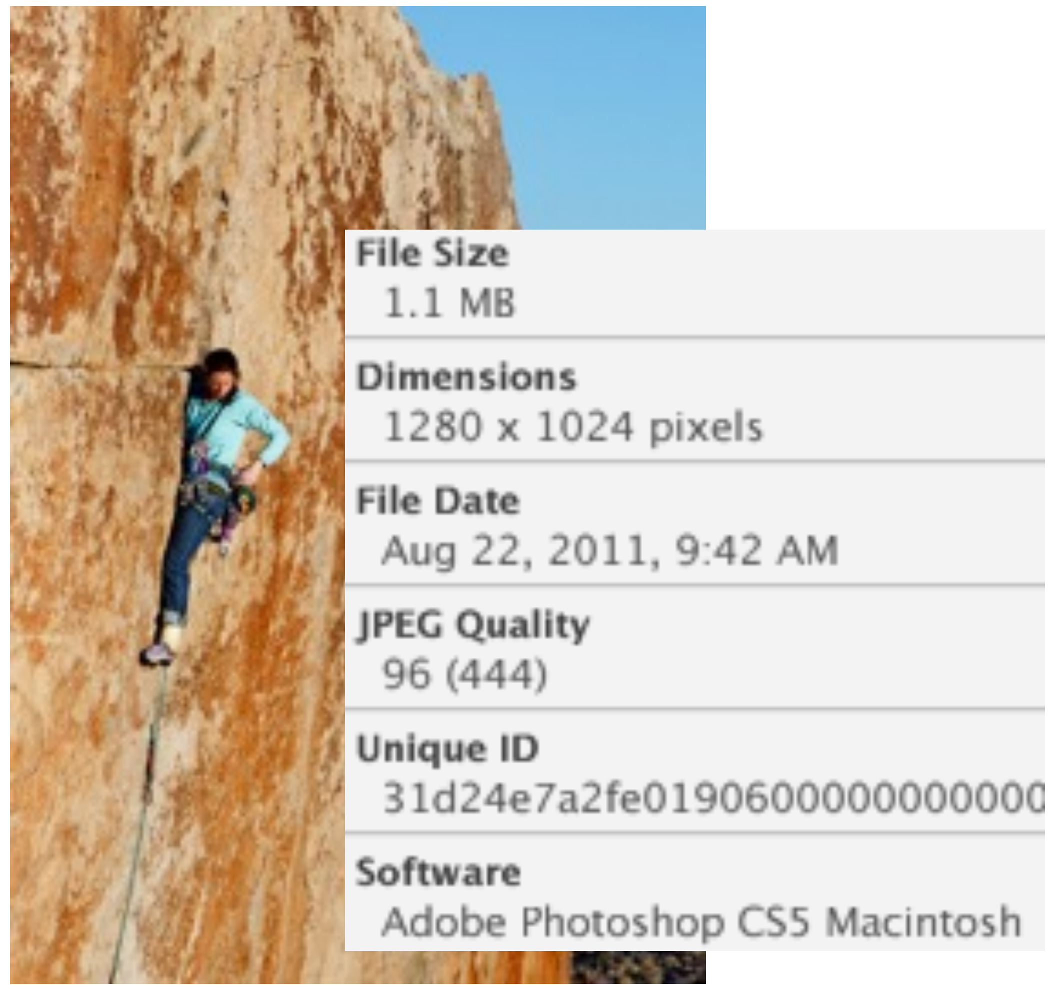
Location
Size
Owner
Group
Permissions
Last-Modified
…

Biomedical



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

Image



File Size
1.1 MB

Dimensions
1280 x 1024 pixels

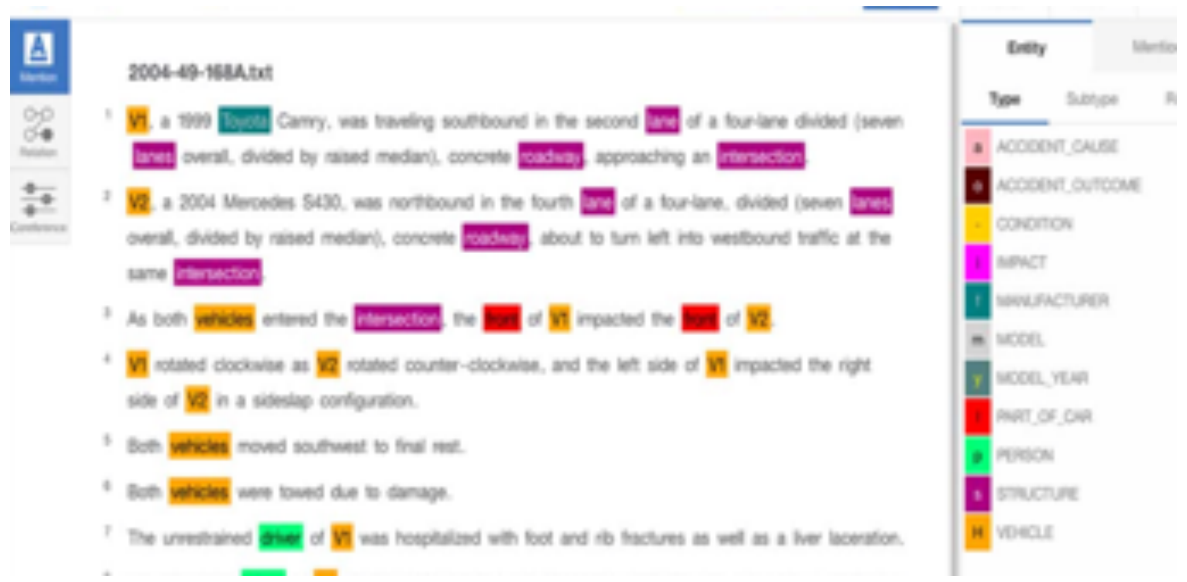
File Date
Aug 22, 2011, 9:42 AM

JPEG Quality
96 (444)

Unique ID
31d24e7a2fe0190600000000000

Software
Adobe Photoshop CS5 Macintosh

Natural Language Processing



2004-49-168A.txt

1 V1, a 1999 Toyota Camry, was traveling southbound in the second lane of a four-lane divided (seven lanes overall, divided by raised median), concrete roadway, approaching an intersection.

2 V2, a 2004 Mercedes S430, was northbound in the fourth lane of a four-lane, divided (seven lanes overall, divided by raised median), concrete roadway, about to turn left into westbound traffic at the same intersection.

3 As both vehicles entered the intersection, the front of V1 impacted the front of V2.

4 V1 rotated clockwise as V2 rotated counter-clockwise, and the left side of V1 impacted the right side of V2 in a sideslap configuration.

5 Both vehicles moved southwest to final rest.

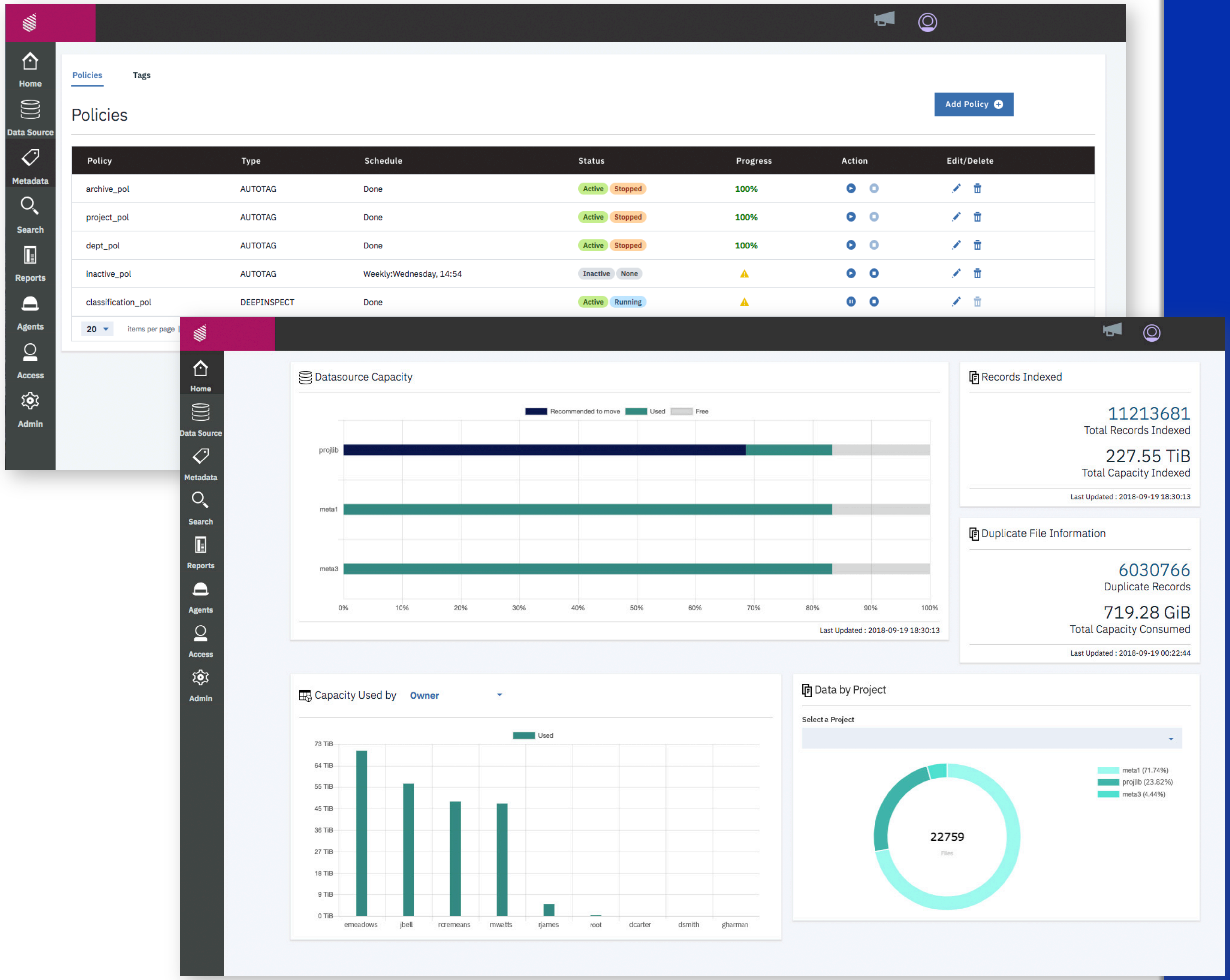
6 Both vehicles were towed due to damage.

7 The unrestrained driver of V1 was hospitalized with foot and rib fractures as well as a liver laceration.

PYTORCH

TensorFlow

IBM Spectrum Discover



Policies

Policy	Type	Schedule	Status	Progress	Action	Edit/Delete
archive_pol	AUTOTAG	Done	Active Stopped	100%		
project_pol	AUTOTAG	Done	Active Stopped	100%		
dept_pol	AUTOTAG	Done	Active Stopped	100%		
inactive_pol	AUTOTAG	Weekly:Wednesday, 14:54	Inactive None	⚠		
classification_pol	DEEPINSPECT	Done	Active Running	⚠		

20 items per page

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

Product Features and Architecture

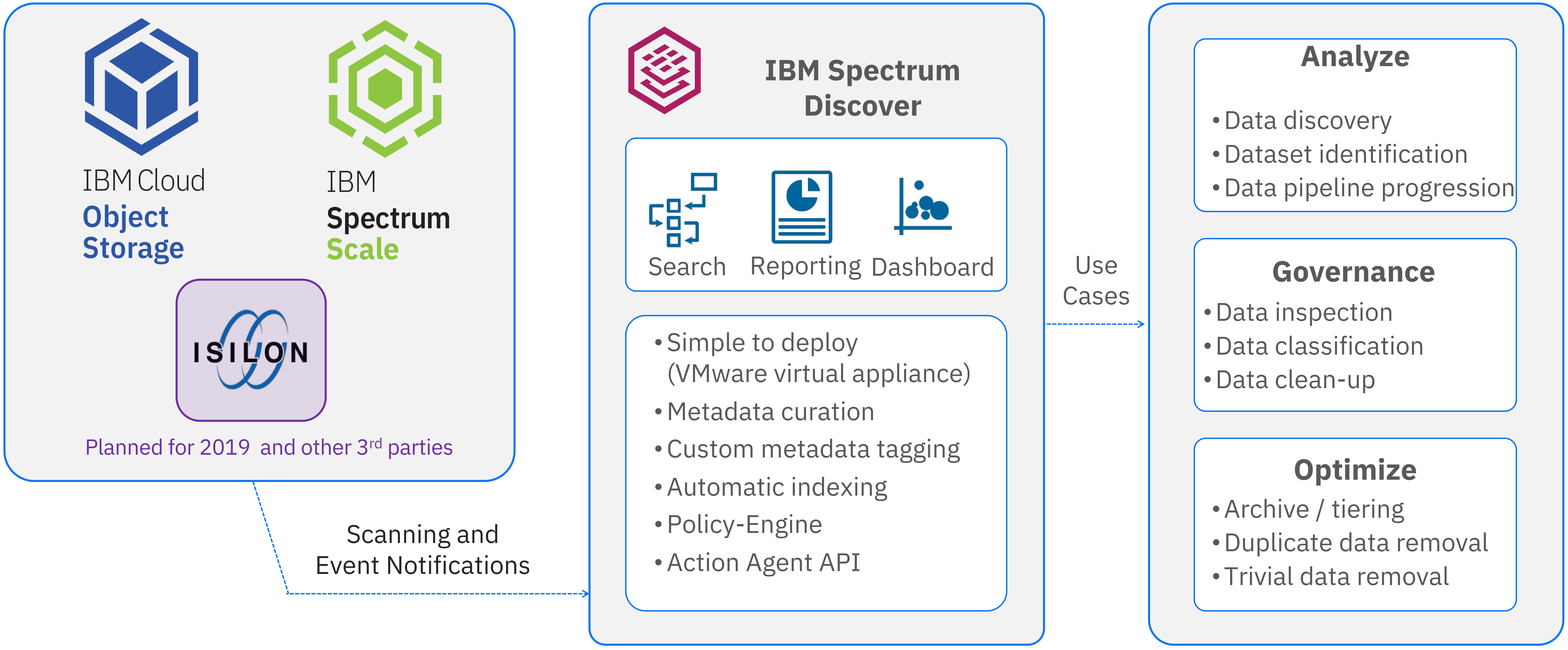
IBM Spectrum Discover Architecture

IBM Spectrum Discover

File and Object Storage

Data Insight

Data Activation/Optimization



IBM Cloud **Object Storage**

IBM **Spectrum Scale**

ISILON

Planned for 2019 and other 3rd parties

Scanning and Event Notifications

IBM Spectrum Discover

Search Reporting Dashboard

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

Use Cases

Analyze

- Data discovery
- Dataset identification
- Data pipeline progression

Governance

- Data inspection
- Data classification
- Data clean-up

Optimize

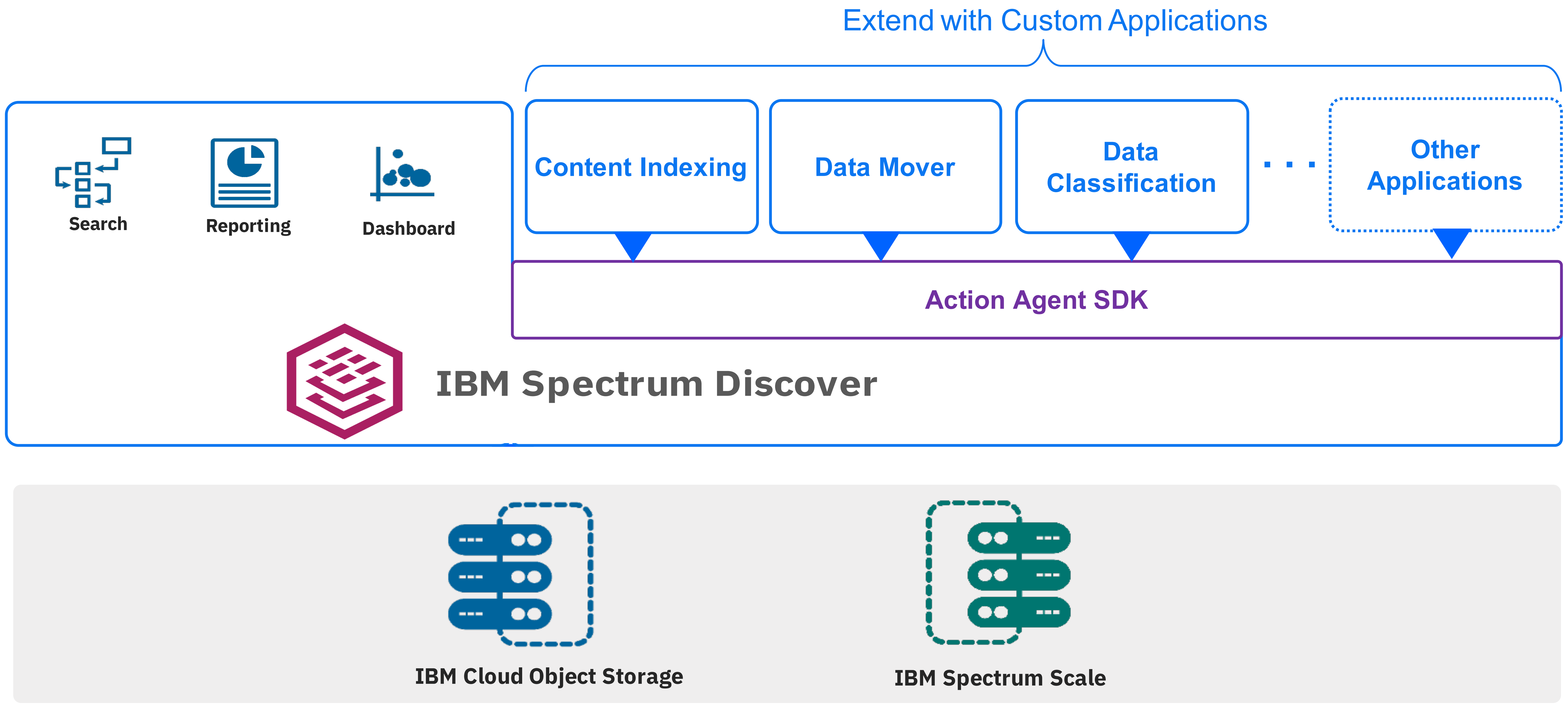
- Archive / tiering
- Duplicate data removal
- Trivial data removal

Extensible Foundation for Data Insight

- Action Agent SDK extends capabilities via well defined API

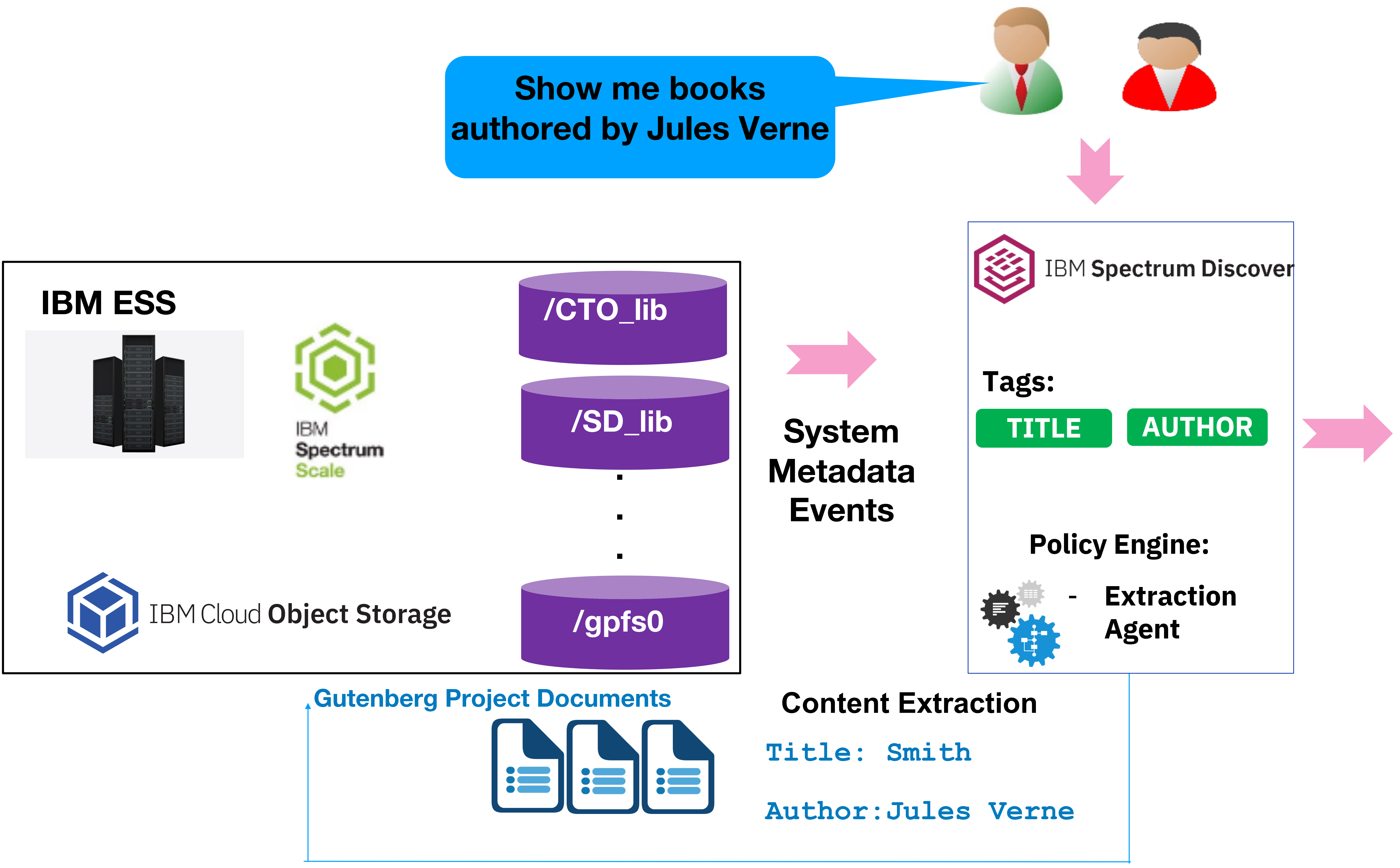
- Customize actions taken based on Discover metadata
  - ❖ Content indexing
  - ❖ Data movement (tiering)
  - ❖ Classification
  - ❖ Sensitive data identification
  - ❖ ROT Detection/Disposal
  - ❖ Etc…

- Integrate with upstream information management applications



Use Case: Identifying relevant data based on content enrichments

IBM Spectrum Discover



Show me books authored by Jules Verne

IBM ESS

IBM Spectrum Scale

/CTO_lib

/SD_lib

. . .

/gpfs0

IBM Cloud Object Storage

System Metadata Events

IBM Spectrum Discover

Tags:

TITLE AUTHOR

Policy Engine:

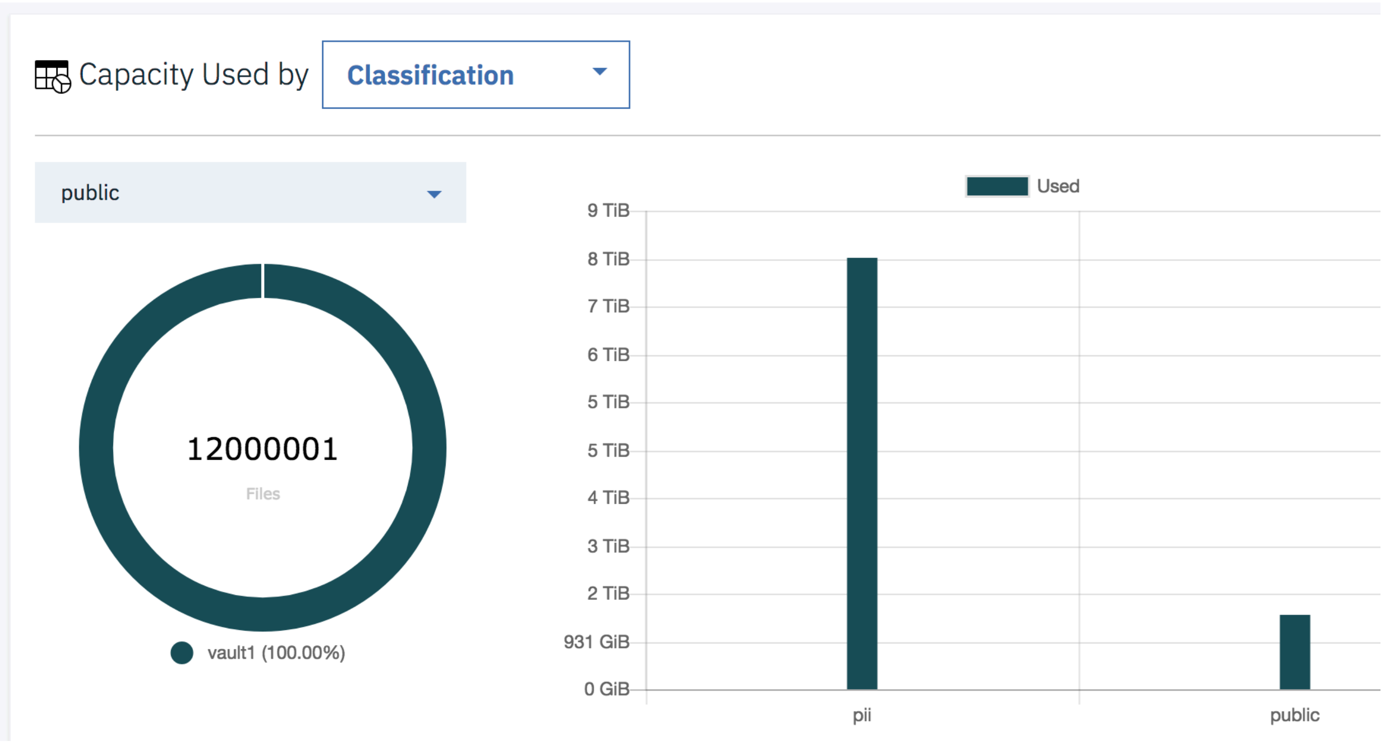
- Extraction Agent

Gutenberg Project Documents

Content Extraction

```
Title:  Smith

Author:Jules Verne
```



Capacity Used by Classification

public

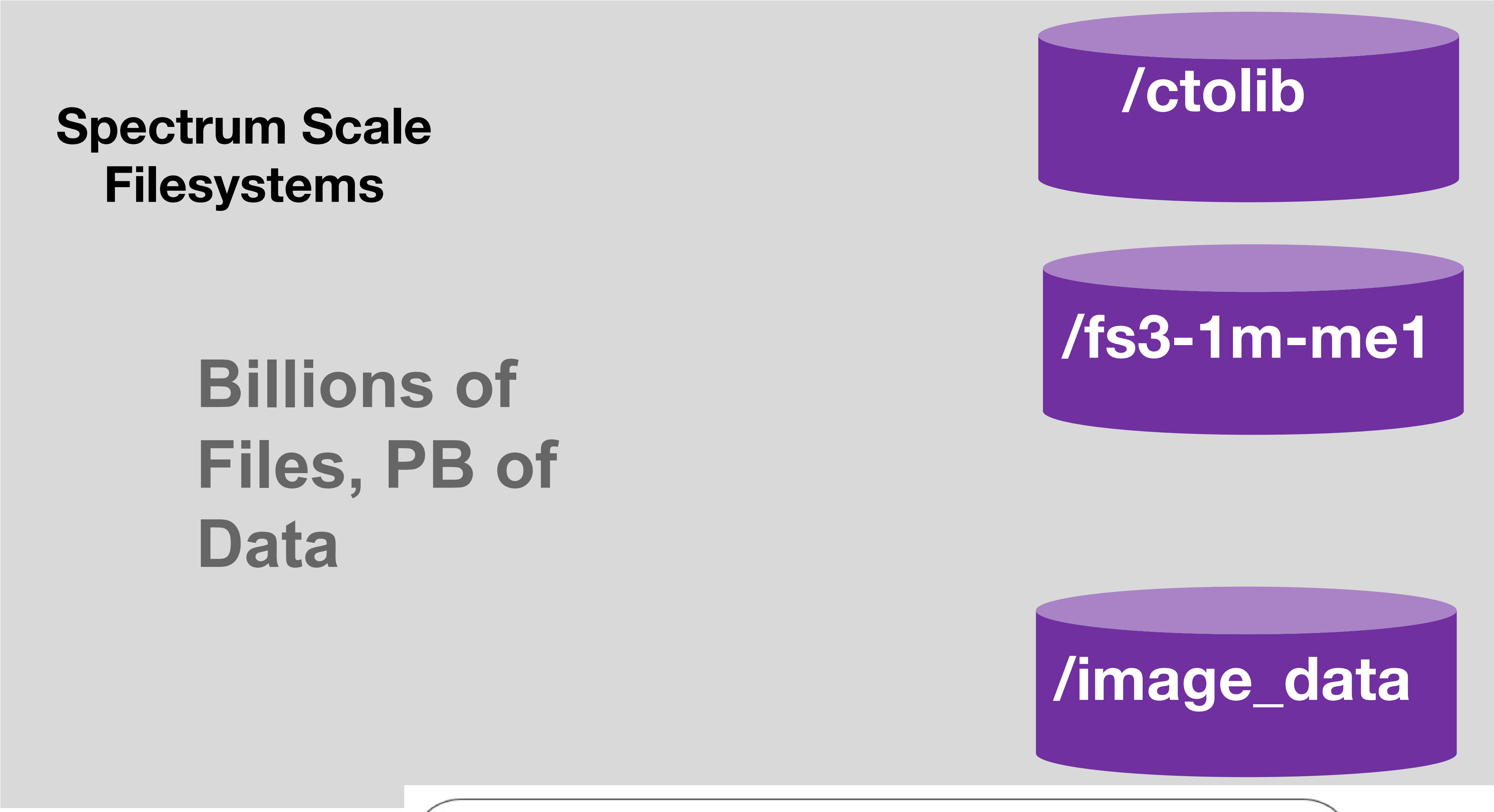
Use Case: Curating the Research Data for Placement Optimization

IBM Spectrum Discover

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

Capacity Reporting

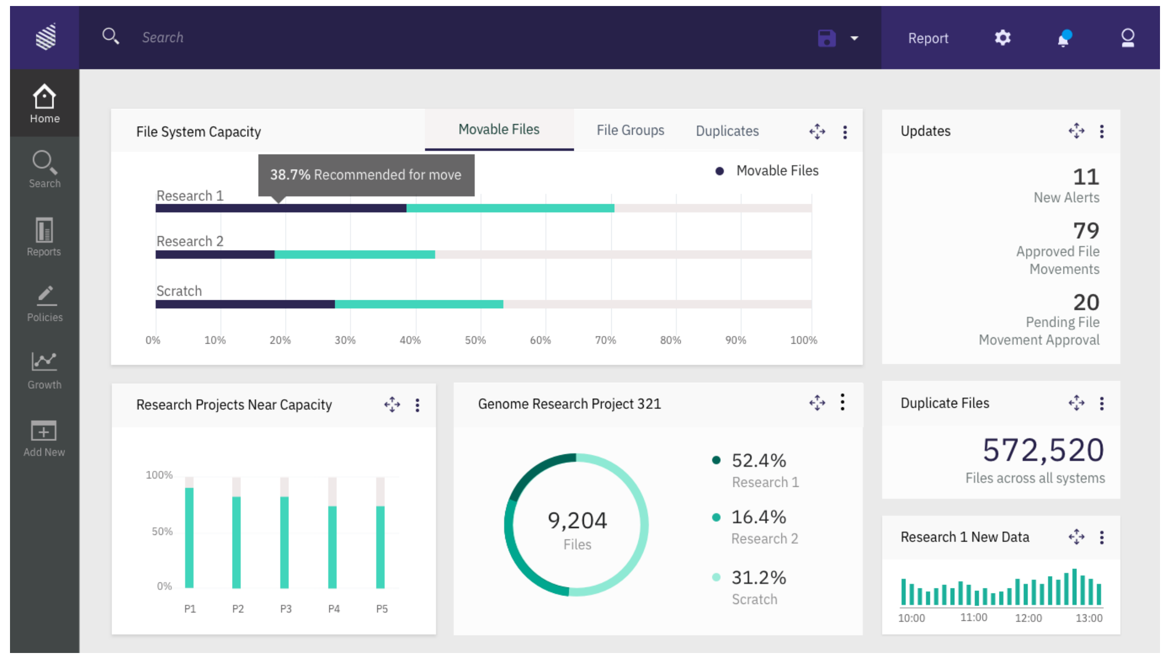
4. Generate reports (Capacity showback)



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

IBM Spectrum Discover

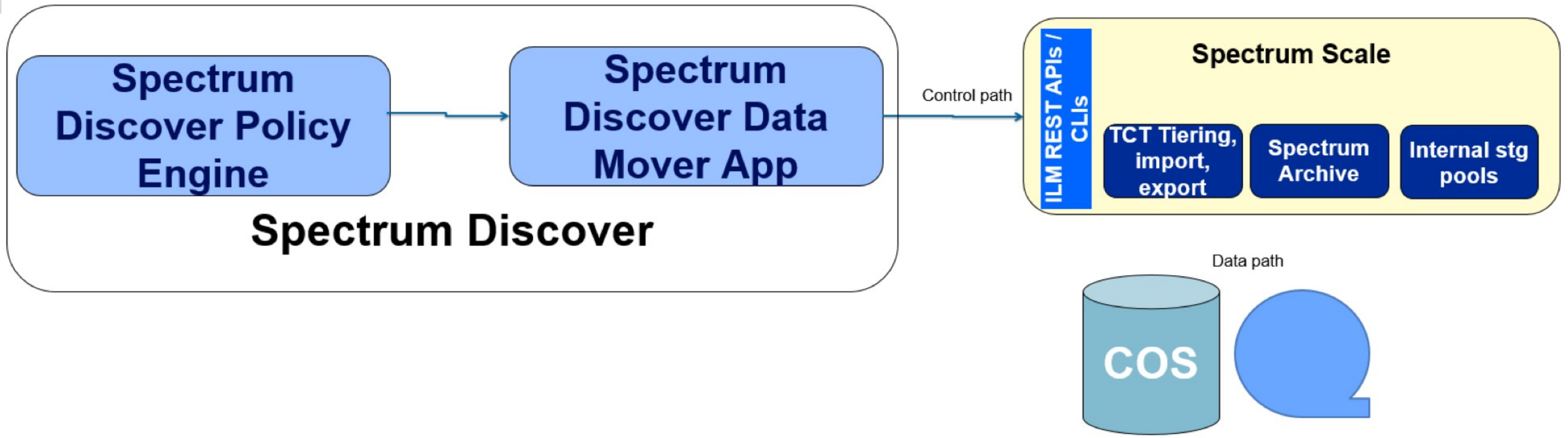
3. Ad-hoc filtered search



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

5. Move to tape

Spectrum Archive



Free Trial Software Download

IBM
Spectrum
Discover

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

THANK YOU!

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

For more information, visit: ibm.com/financing

IBM Cloud – Middleware, New Orchard Road Armonk, NY 10504.
Produced in the United States of America, August 2017.