IBM **Spectrum Discover**

Data Insight for Petabyte-Scale Unstructured Data Storage

Indulis Bernsteins
Systems & Storage Architect

CIUK, Manchester
December 12, 2018

Please note

IBM’s statements regarding its plans, directions, and intent are subject to change or withdrawal without notice and at IBM’s sole discretion.

Information regarding potential future products is intended to outline our general product direction and it should not be relied on in making a purchasing decision.

The information mentioned regarding potential future products is not a commitment, promise, or legal obligation to deliver any material, code or functionality. Information about potential future products may not be incorporated into any contract.

The development, release, and timing of any future features or functionality described for our products remains at our sole discretion.

Performance is based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput or performance that any user will experience will vary depending upon many factors, including considerations such as the amount of multiprogramming in the user’s job stream, the I/O configuration, the storage configuration, and the workload processed. Therefore, no assurance can be given that an individual user will achieve results similar to those stated here.

Harnessing the Value of Data

> “The world’s most valuable resource is no longer oil, but **data**.”
>
> *The Economist, May, 2017*

*…how to **harness the value?***

- **Identify**
- **Categorize**
- **Utilize**

Unstructured Data Challenges

Enterprises with

1,000 TB+

unstructured data

grew 3X

2016 ⇨ 2017

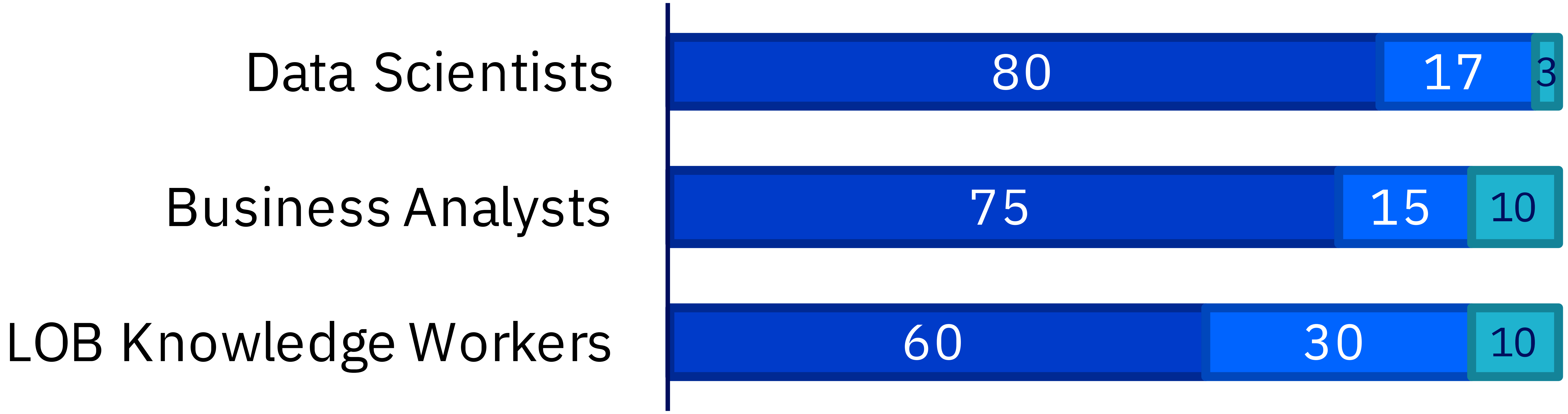
39%

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 Challenges

Data Scientists spend $^{2}/_{3}$ of their time finding data



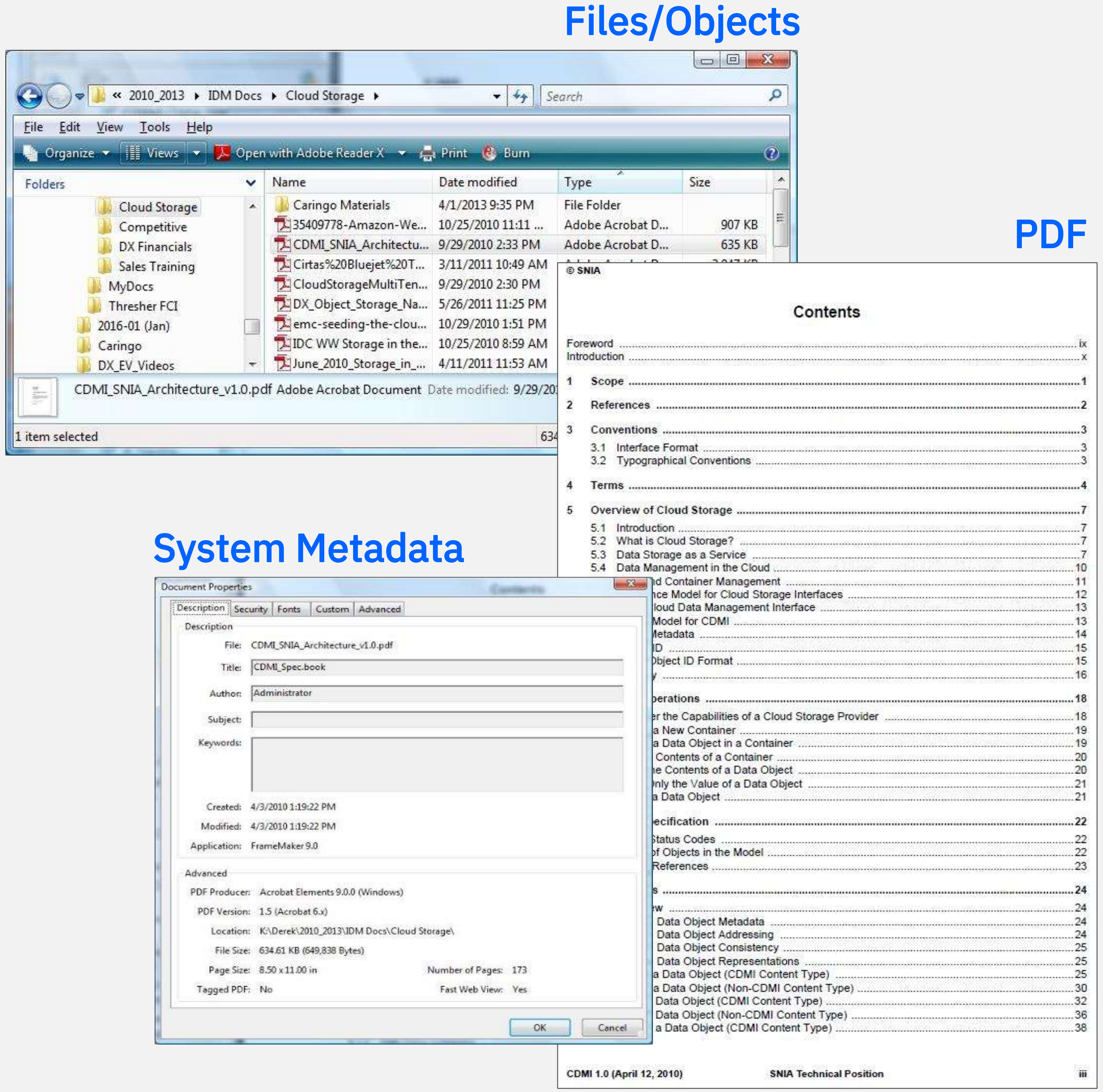
Challenges for exabyte-scale data: how to…

- Pinpoint & activate relevant data for AI, analytics etc.
- Get fine-grained visibility of value of data
- Remove redundant, trivial & obsolete (“ROT”) data
- Identify & classify sensitive data

Metadata is data about data

Provides **context** to **classify** & **manage** unstructured data

- Last time accessed
- Author/user
- Type of Data
- Key words
- Custom tags
  - e.g. project, department, etc.



Name	Date modified	Type	Size
Caringo Materials	4/1/2013 9:35 PM	File Folder	
35409778-Amazon-We...	10/25/2010 11:11 ...	Adobe Acrobat D...	907 KB
CDMI_SNIA_Architectu...	9/29/2010 2:33 PM	Adobe Acrobat D...	635 KB
Cirtas%20Bluejet%20T...	3/11/2011 10:49 AM		
CloudStorageMultiTen...	9/29/2010 2:30 PM		
DX_Object_Storage_Na...	5/26/2011 11:25 PM		
emc-seeding-the-clou...	10/29/2010 1:51 PM		
IDC WW Storage in the...	10/25/2010 8:59 AM		
June_2010_Storage_in_...	4/11/2011 11:53 AM		

CDMI_SNIA_Architecture_v1.0.pdf Adobe Acrobat Document Date modified: 9/29/20

© SNIA

Contents

Foreword ix
Introduction x

1 Scope 1
2 References 2
3 Conventions 3
3.1 Interface Format 3
3.2 Typographical Conventions 3
4 Terms 4
5 Overview of Cloud Storage 7
5.1 Introduction 7
5.2 What is Cloud Storage? 7
5.3 Data Storage as a Service 7
5.4 Data Management in the Cloud 10
d Container Management 11
nce Model for Cloud Storage Interfaces 12
loud Data Management Interface 13
Model for CDMI 13
letadata 14
D 15
bject ID Format 15
y 16
perations 18
r the Capabilities of a Cloud Storage Provider 18
a New Container 19
a Data Object in a Container 19
Contents of a Container 20
e Contents of a Data Object 20
nly the Value of a Data Object 21
a Data Object 21
ecification 22
Status Codes 22
f Objects in the Model 22
References 23
s 24
w 24
Data Object Metadata 24
Data Object Addressing 24
Data Object Consistency 25
Data Object Representations 25
a Data Object (CDMI Content Type) 25
a Data Object (Non-CDMI Content Type) 30
Data Object (CDMI Content Type) 32
Data Object (Non-CDMI Content Type) 36
a Data Object (CDMI Content Type) 38

CDMI 1.0 (April 12, 2010) SNIA Technical Position iii

Document Properties

Description

File: CDMI_SNIA_Architecture_v1.0.pdf
Title: CDMI_Spec.book
Author: Administrator
Subject:
Keywords:
Created: 4/3/2010 1:19:22 PM
Modified: 4/3/2010 1:19:22 PM
Application: FrameMaker 9.0

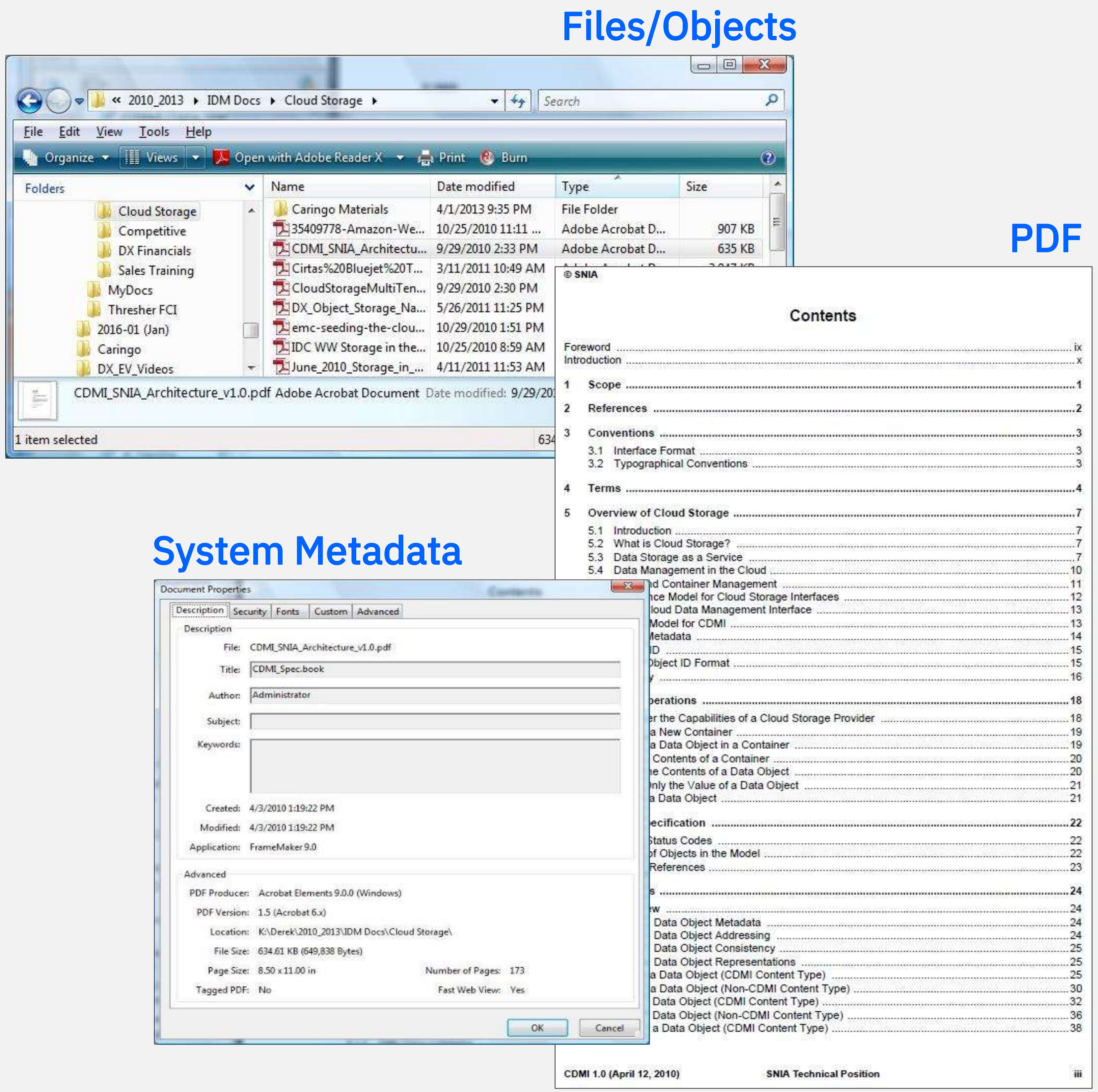
Advanced

PDF Producer: Acrobat Elements 9.0.0 (Windows)
PDF Version: 1.5 (Acrobat 6.x)
Location: K:\Derek\2010_2013\IDM Docs\Cloud Storage\
File Size: 634.61 KB (649,838 Bytes)
Page Size: 8.50 x 11.00 in Number of Pages: 173
Tagged PDF: No Fast Web View: Yes

Types of metadata

System: Typical file system metadata, date created, owner, last modified, size, file type, etc.

Custom: User-defined or based on unique organizational schema/taxonomy



Files/Objects

« 2010_2013 ▸ IDM Docs ▸ Cloud Storage ▸
Search
File Edit View Tools Help
Organize Views Open with Adobe Reader X Print Burn
Folders
Cloud Storage
Competitive
DX Financials
Sales Training
MyDocs
Thresher FCI
2016-01 (Jan)
Caringo
DX_EV_Videos

Name	Date modified	Type	Size
Caringo Materials	4/1/2013 9:35 PM	File Folder	
35409778-Amazon-We...	10/25/2010 11:11 ...	Adobe Acrobat D...	907 KB
CDMI_SNIA_Architectu...	9/29/2010 2:33 PM	Adobe Acrobat D...	635 KB
Cirtas%20Bluejet%20T...	3/11/2011 10:49 AM		
CloudStorageMultiTen...	9/29/2010 2:30 PM		
DX_Object_Storage_Na...	5/26/2011 11:25 PM		
emc-seeding-the-clou...	10/29/2010 1:51 PM		
IDC WW Storage in the...	10/25/2010 8:59 AM		
June_2010_Storage_in_...	4/11/2011 11:53 AM		

CDMI_SNIA_Architecture_v1.0.pdf Adobe Acrobat Document Date modified: 9/29/20

1 item selected 634

PDF

© SNIA

Contents

CDMI 1.0 (April 12, 2010) SNIA Technical Position iii

System Metadata

Document Properties

Description

File: CDMI_SNIA_Architecture_v1.0.pdf
Title: CDMI_Spec.book
Author: Administrator
Subject:
Keywords:

Created: 4/3/2010 1:19:22 PM
Modified: 4/3/2010 1:19:22 PM
Application: FrameMaker 9.0

Advanced

PDF Producer: Acrobat Elements 9.0.0 (Windows)
PDF Version: 1.5 (Acrobat 6.x)
Location: K:\Derek\2010_2013\IDM Docs\Cloud Storage\
File Size: 634.61 KB (649,838 Bytes)
Page Size: 8.50 x 11.00 in Number of Pages: 173
Tagged PDF: No Fast Web View: Yes

System Metadata Collected by Spectrum Discover

IBM Spectrum Scale

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

- Install policy scan agent on node in target Spectrum Scale cluster
- No EAs/XATTRs
- No user MD from Objects

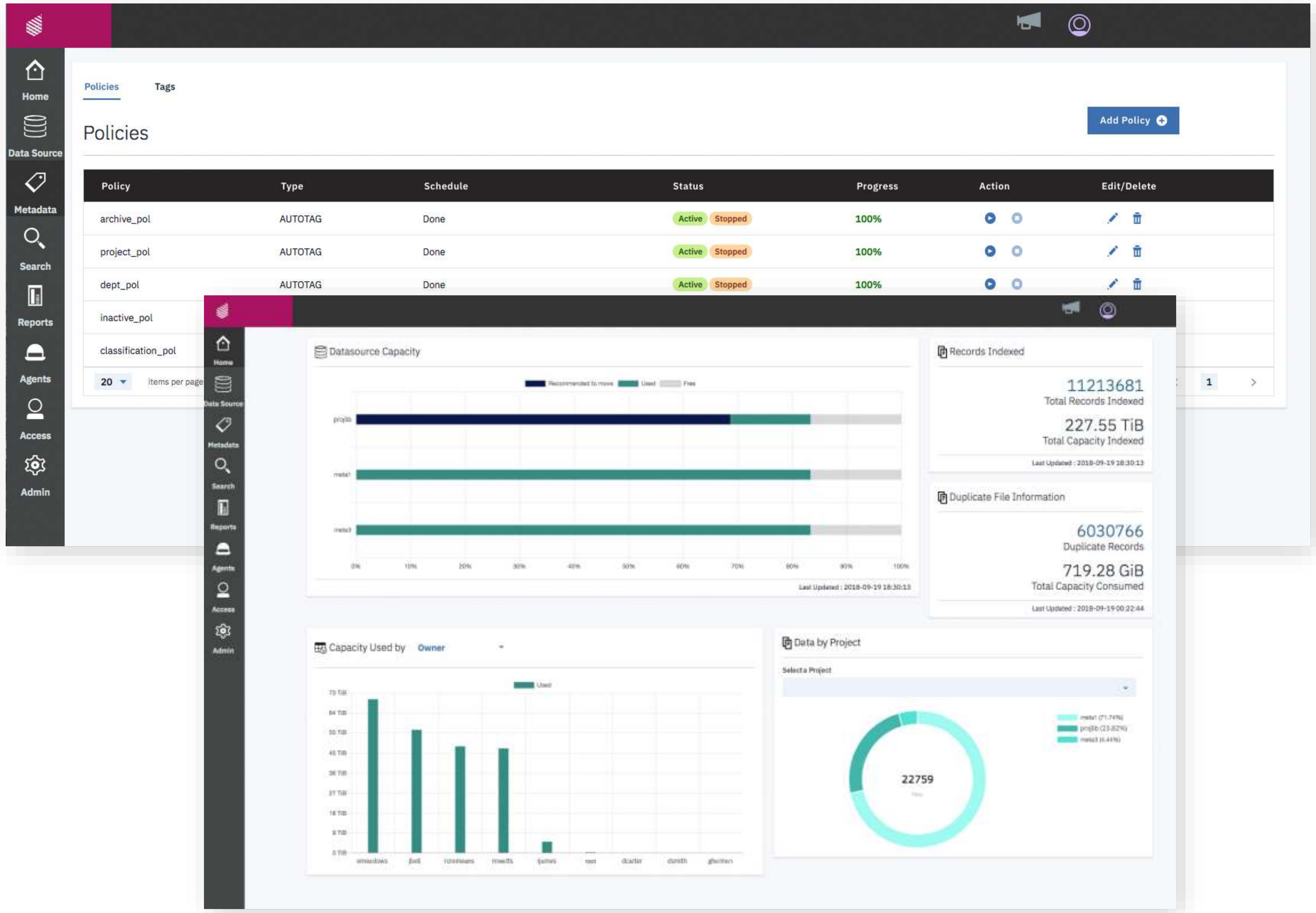
IBM Cloud Object Storage

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

IBM **Spectrum Discover**

Data Insight for Analytics, Governance & Optimization

- Automated cataloging
- Comprehensive insights: combine system metadata with custom tags
- Extend usefulness using the API, custom tags, content inspection and policy-based workflows



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						
classification_pol						

20 items per page 1

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

Spectrum Discover **is**

- A scalable & simple to use virtualised s/w “appliance”
  - Capable of dealing with System & Custom metadata
  - Works with Billions of files & objects
  - Designed to be easy to use & support admins **and** end users

- …Uses System and Custom Metadata to primarily support **storage management**
  - Extensible: add MD tags from “deep inspections”
  - For file and object curation: movement, deletion, migration
  - For data selection for processing: ad-hoc selection

IBM **Spectrum Discover**

Spectrum Discover **is not**

- Not a replacement for all Metadata engines
  - e.g. DICOM scanner to populate hospital patient record system

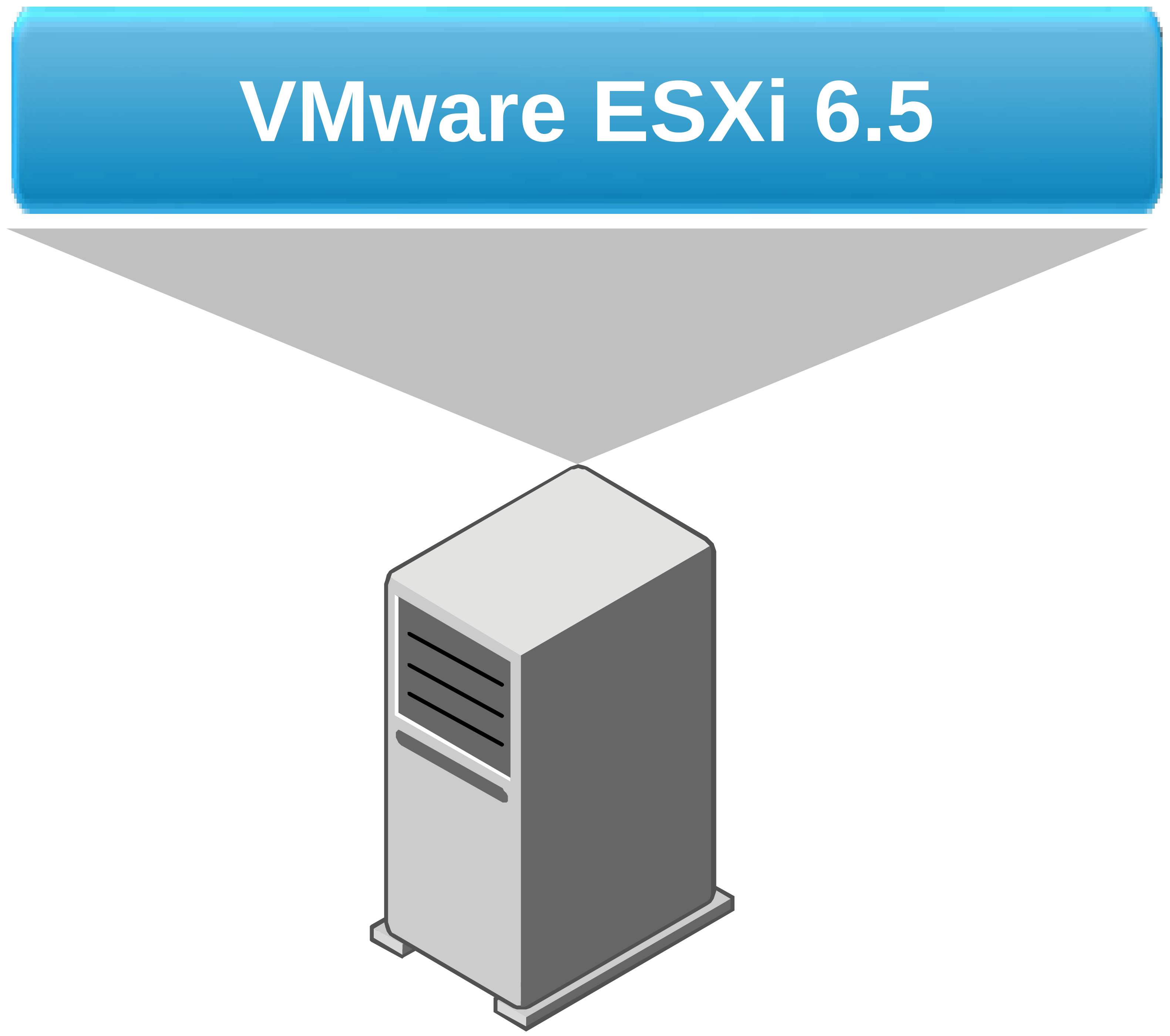
- Not a replacement for all ILM/HSM features in Spectrum Scale
  - There is some overlap, as Data Migration is planned (SoD)
    - Use cases for Spectrum Discover are ***probably*** more ad-hoc, user driven

IBM **Spectrum Discover**

Single Node Spectrum Discover Deployment Example (2 billion files)

Spectrum Discover

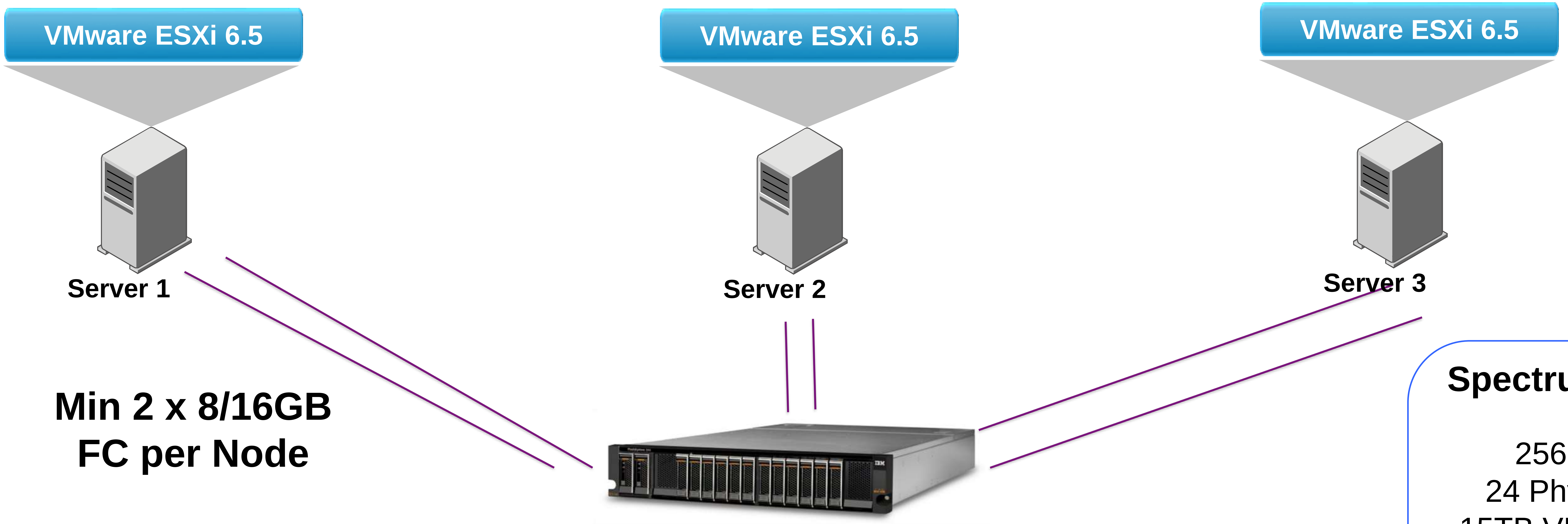
128GB RAM
8 Physical CPU
5TB VMDK primary
2TB VMDK backup
1/10GbE Virtual Network



VMware ESXi 6.5

Local attached SSD or shared storage

Multi-Node Spectrum Discover Deployment Example (10 billion files)



Min 2 x 8/16GB FC per Node

Optional FC Switch

Shared storage (Flash/SSD/NVMe) e.g. IBM Flashsystem

Spectrum Discover

256GB RAM
24 Physical CPU
15TB VMDK primary
7TB VMDK backup
10GbE Virtual Network

Use Cases

Optimization Use Case:

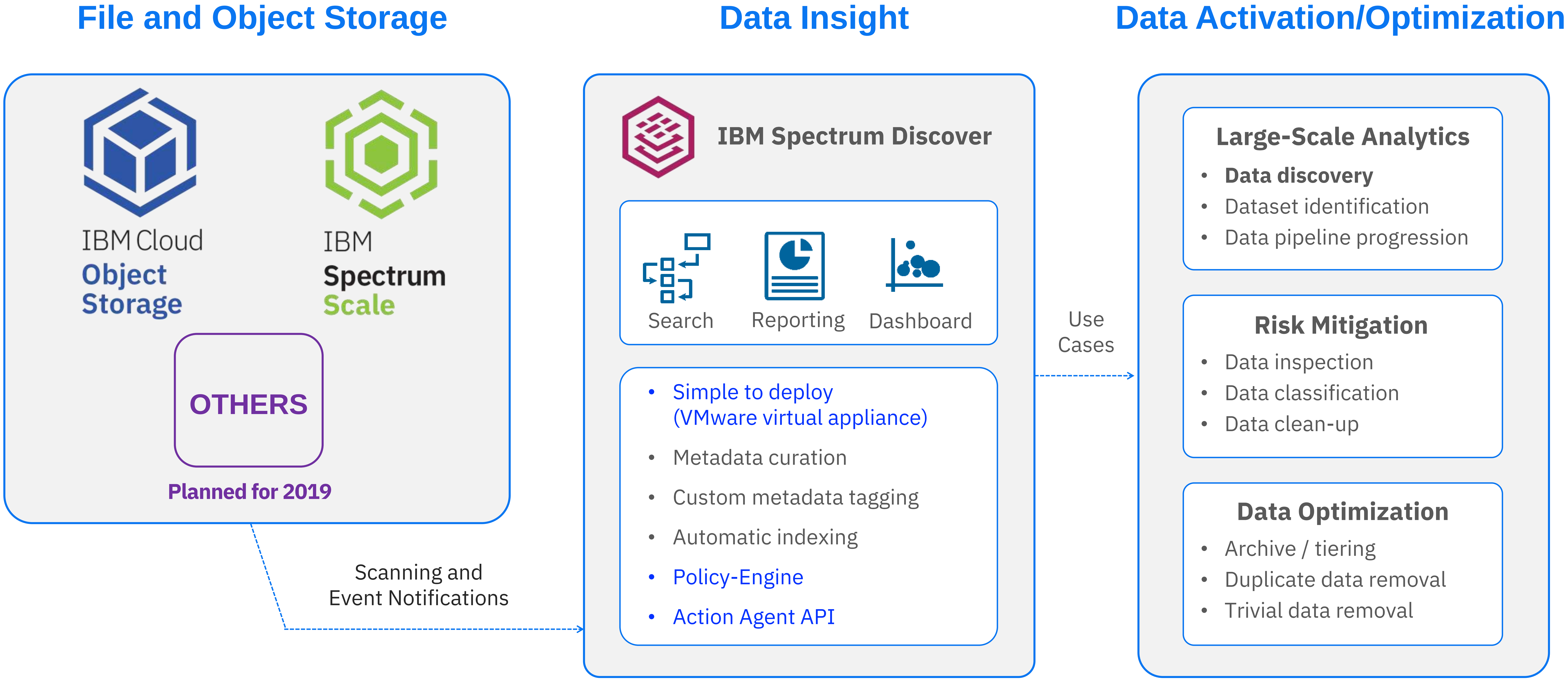
Optimize Data Placement on Storage

IBM Spectrum Discover

		Problem	Solution
Business Leaders (Dir / VP of IT / CIO / CFO) Cost • Unnecessary purchases due to inefficient management of storage • Avoid unplanned Operational Expense due to business processes not being followed Risk • Ensuring new Business Strategies are not delayed by lack of capacity • Ensuring existing applications have capacity for expansion	**IT Infrastructure Team** (Storage Administrator)	**Managing Capacity Utilization** • Monitor multiple vendor technologies • Wait for the signs of quota threshold violations • Track down Data Owners & ask them to audit their data placement, retention, & take corrective action • Wait for reply from Data Owner regarding what data can be archived	**Shared Management Responsibility** • Visualize/monitor utilization of multiple storage sources with drill-down analytics. • Custom, policy-driven data tagging identifies candidate files for action. • Intelligently manage data placement on appropriate storage tier (archive)
	Data Owners (Researcher / Data Scientist / Lab Manager / RIM)	**Manually Managing the Data** • Track data ownership in a spreadsheet • Relies on peers following best-practices in order to track data • Tries to figure out what data can be archived in order to save on storage costs • Open ticket to have data archived & migrated	**Shared Management Responsibility** • Custom, policy-driven data tagging identifies candidate files for action • Multi-faceted search to identify other candidates for archiving. • Improved understanding of data use profile

Product Features & Architecture

IBM Spectrum Discover Overview



File and Object Storage

IBM Cloud **Object Storage**

IBM **Spectrum Scale**

OTHERS

Planned for 2019

Scanning and Event Notifications

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

Use Cases

Data Activation/Optimization

Large-Scale Analytics

- **Data discovery**
- Dataset identification
- Data pipeline progression

Risk Mitigation

- Data inspection
- Data classification
- Data clean-up

Data Optimization

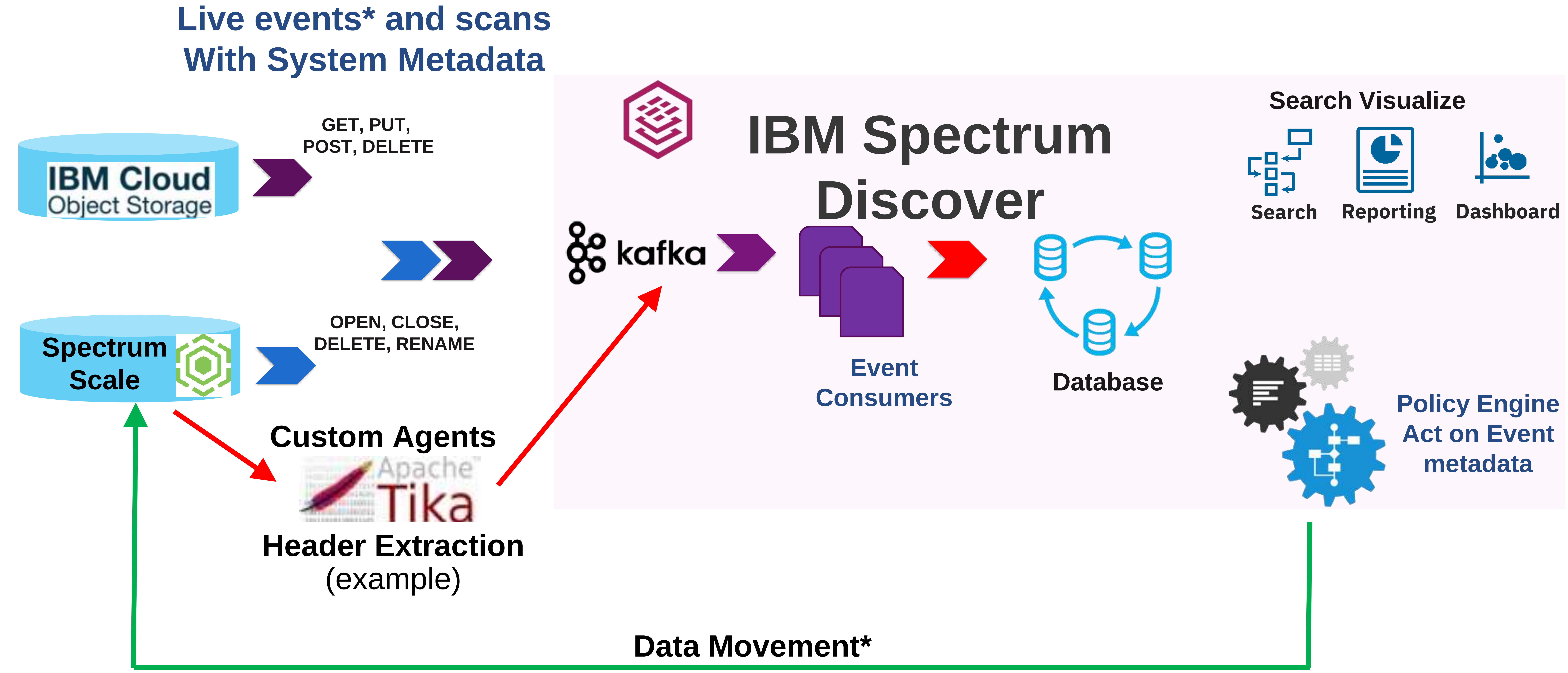
- Archive / tiering
- Duplicate data removal
- Trivial data removal

Extensible Foundation for Data Insight

IBM **Spectrum Discover**

Connectors	• Scanner for IBM Cloud Object Storage (COS) and IBM Spectrum Scale • Notifications for IBM Cloud Object Storage • Tech preview of Notifications (Live Events) for IBM Spectrum Scale		
Platform	• **Support for Single and Multi-Node (3-Node HA Cluster) (x86 only)** • Code upgrade for Single and Multi-node • **Encryption of Metadata Database and Notification Logs (Kafka)**	• **Support for Role Based Access Control (RBAC)** • Remote Support – tools to collect logs and upload to IBM support • **Backup / restore / DR of Metadata Database**	• Audit Trail – track and log user actions on Spectrum Discover • Dashboards to monitor Spectrum Discover health
Action Agent SDK Ecosystem	• SDK to help extend the platform capabilities to perform custom actions around data – data migration, archiving, content-based search & tagging, etc.		
GUI	• Basic Search • **Advanced drill-down search** • **Dashboard** to visualize storage consumption on a wide range of system and custom metadata • **Create and schedule policies**	• Support for Role Based Access Control (RBAC)	
Scalability	• Up to 100 billion indexed documents		
Performance	• Ingest up over 1 Billion records per day		
Quality	• Net Promoter Score (NPS) widget instrumented into Discover to gather NPS scores.		

IBM Spectrum Discover Architecture



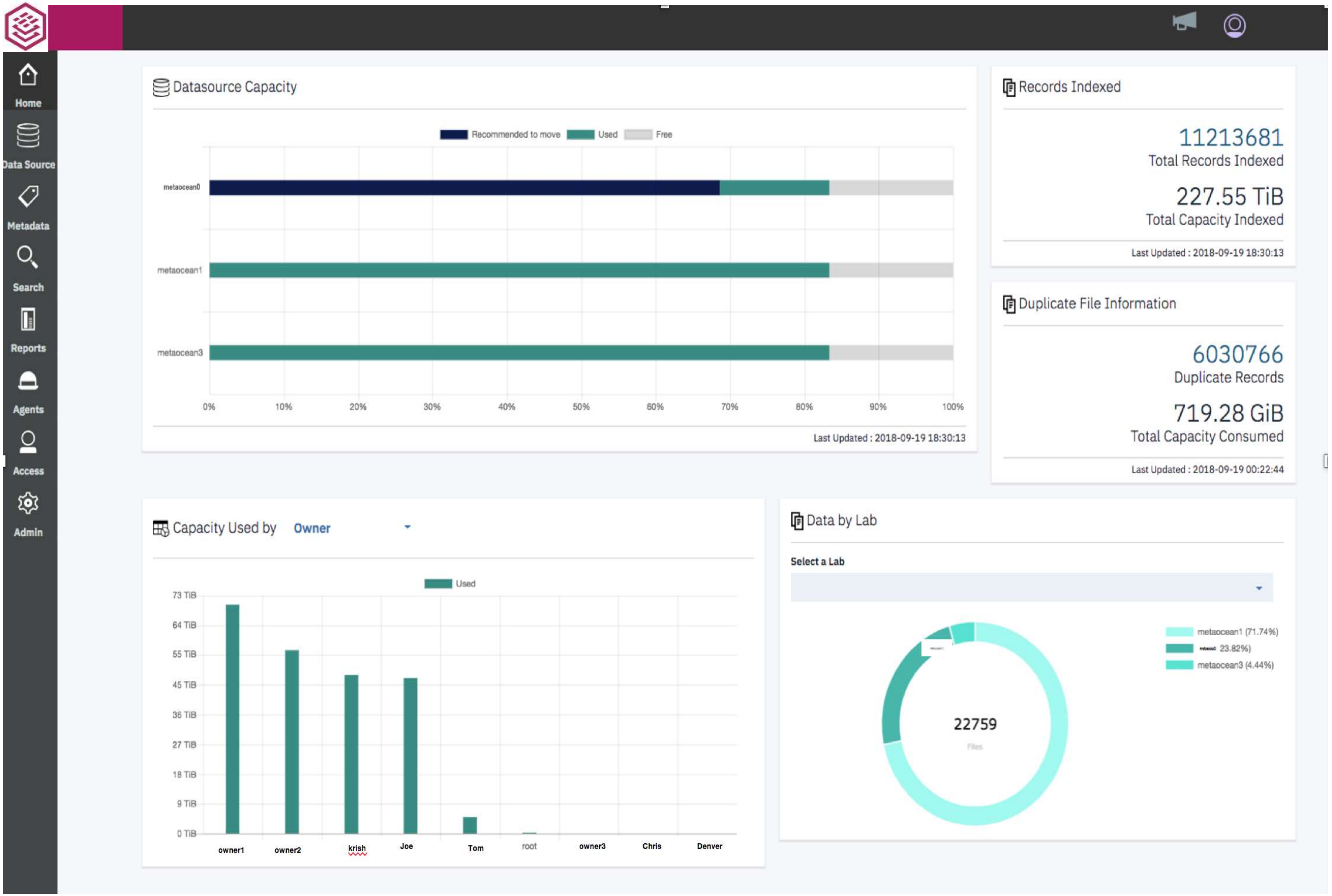
*Planned for 2019

Spectrum Discover Admin Dashboard

IBM
Spectrum
Discover

Monitor storage utilization and data recommendations (Move/Archive)

Preview capacity use by data facet
- Classification
- Owner
- File Type
- Etc.



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 Lab

Select a Lab

Total indexed data and capacity

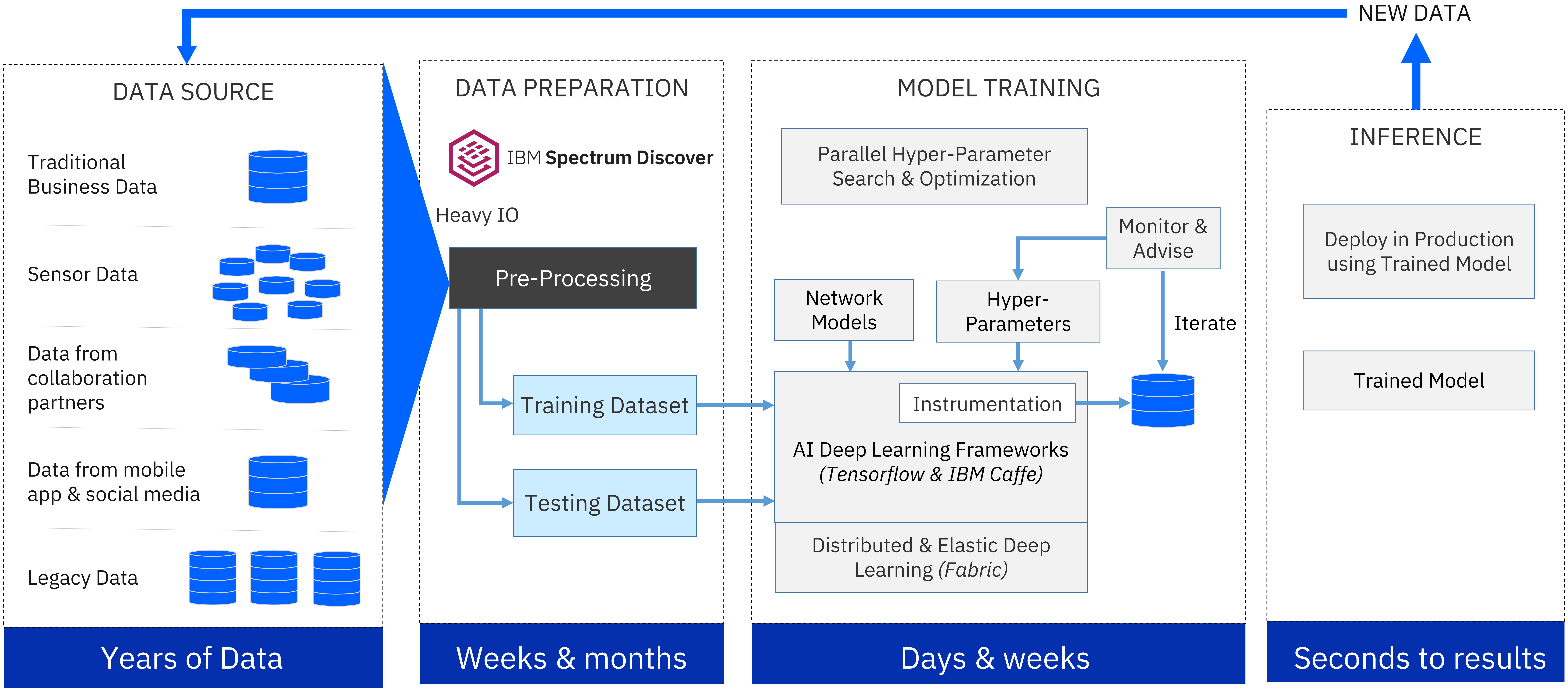
Duplicate file or object candidates
- Number
- Capacity used

Data capacity by group/collection
- Customer defined
- Lab/Project/etc.

owner1 owner2 krish Joe Tom owner3 Chris Denver

Data Curation for AI Workloads

Spectrum Discover for Data Preparation & Data Curation

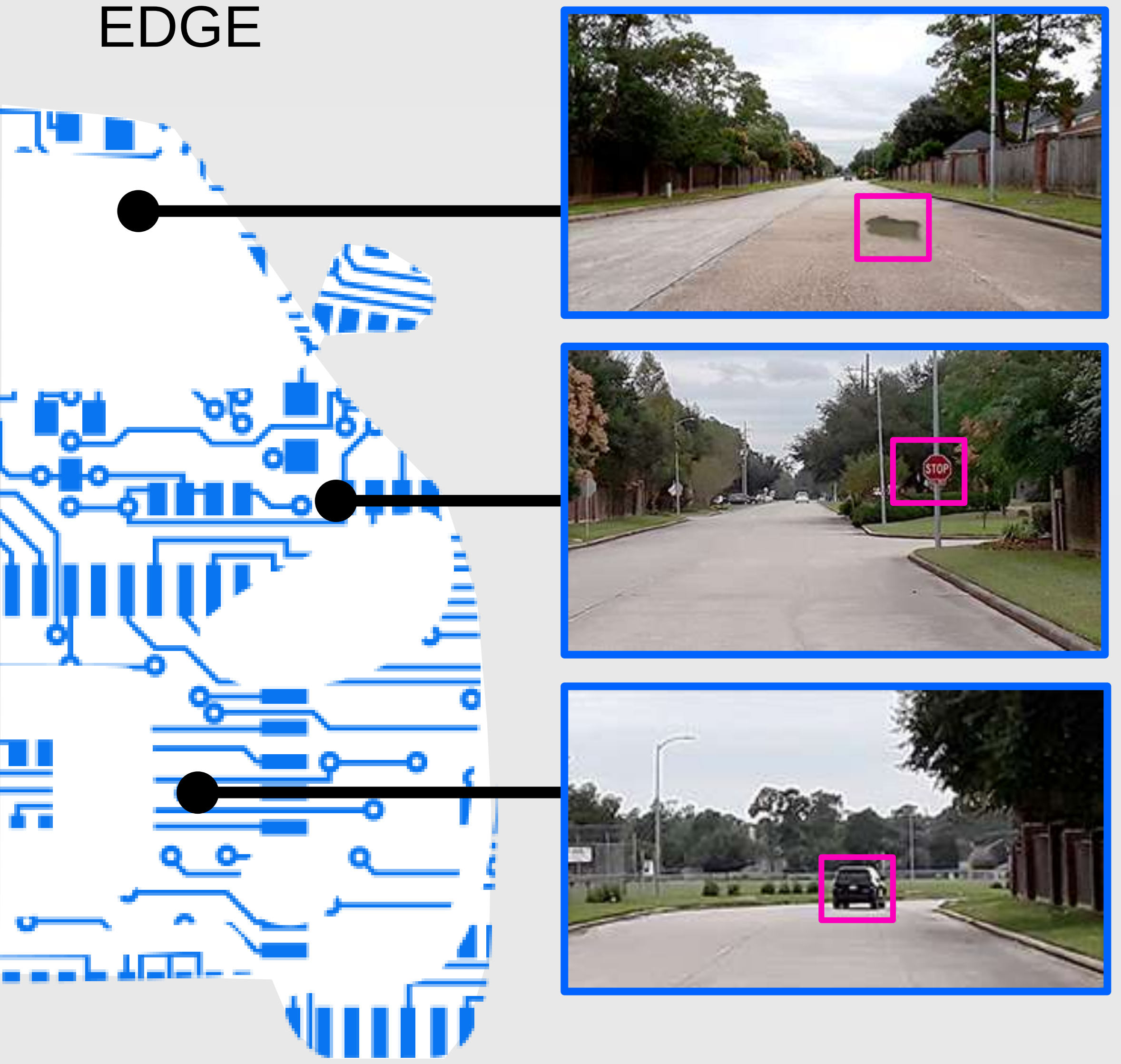


ADAS-AD Data Pipeline

INGEST
CLASSIFY
ANALYZE / TRAIN
INSIGHTS

Autonomous Vehicle Use Case and Demo with Spectrum Discover

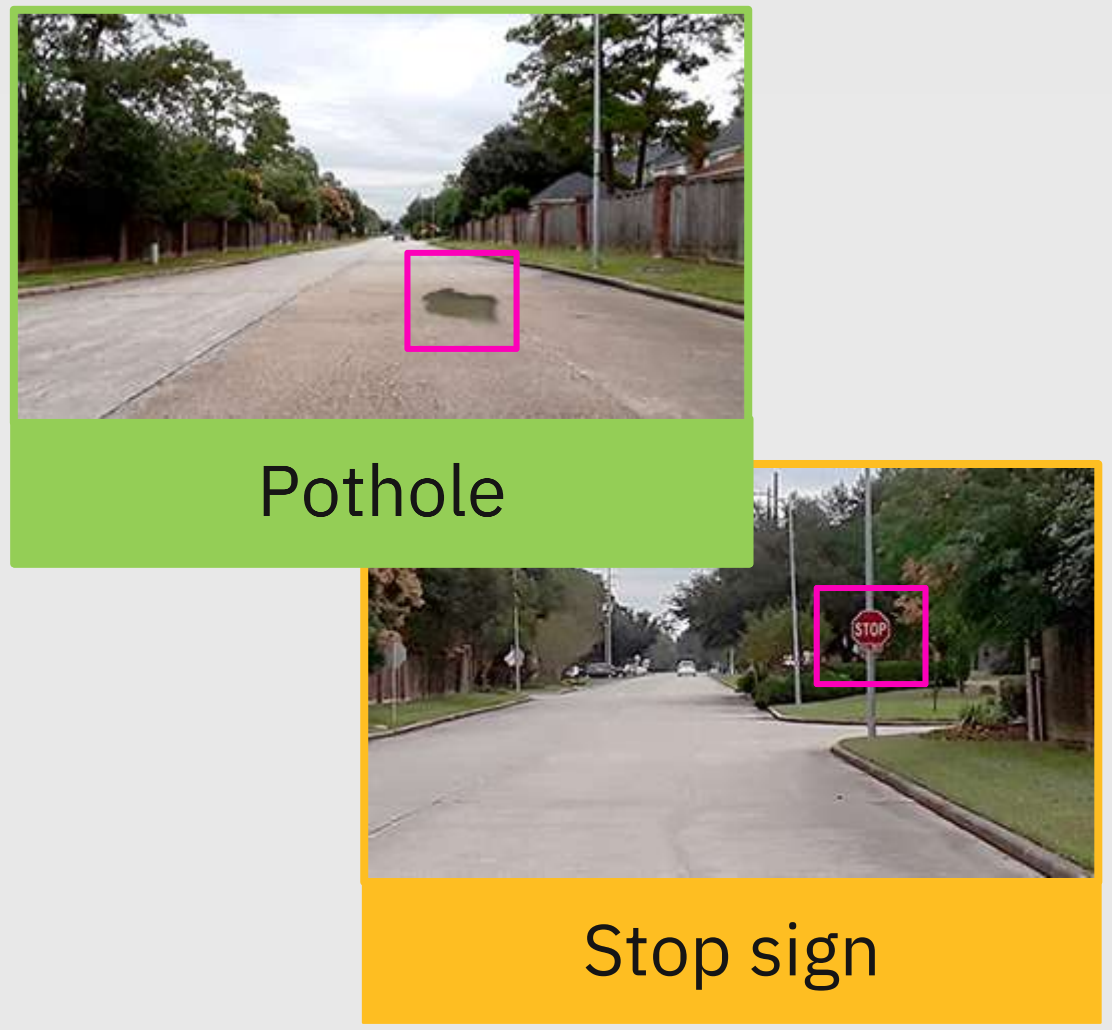
INGEST



IBM Cloud **Object Storage**

- Global ingest of IOT data from vehicles
- Geo-dispersed COS

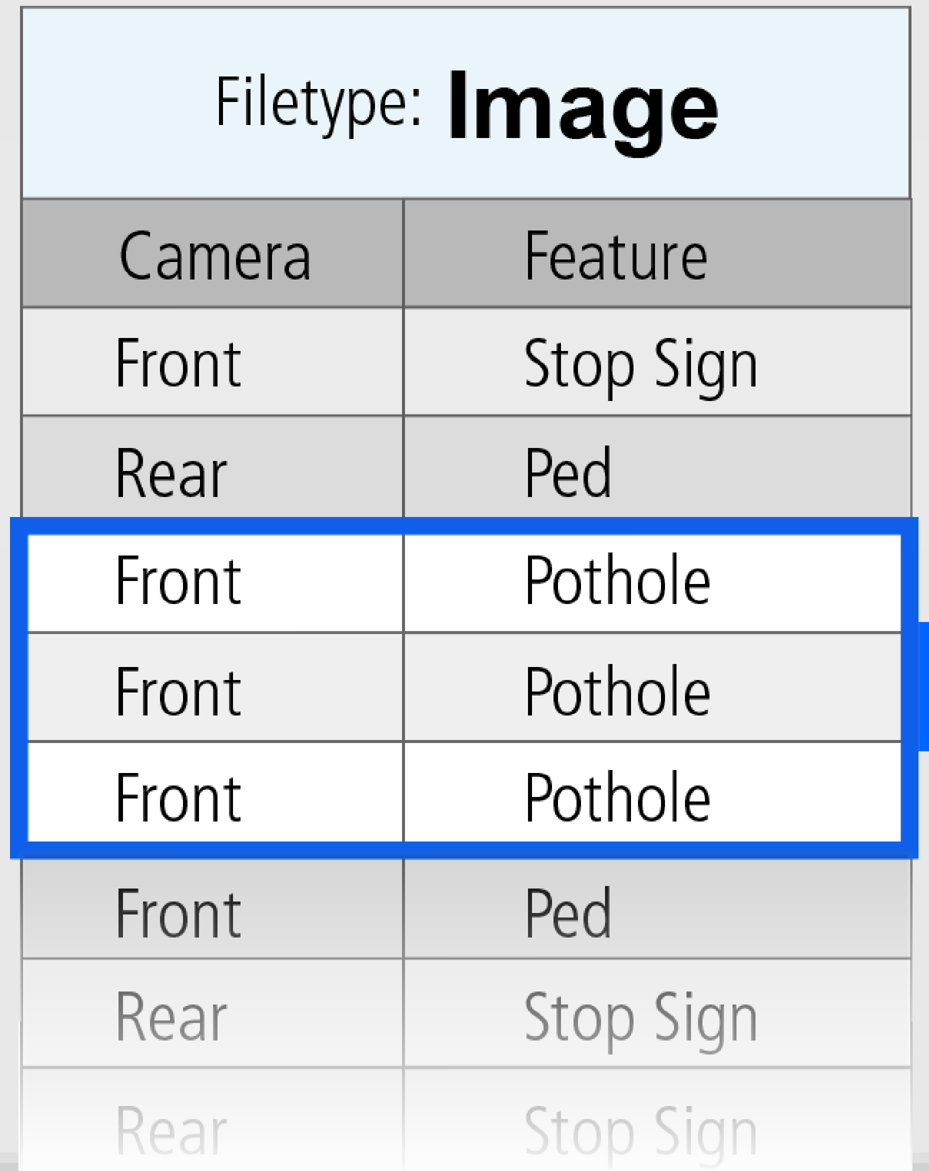
SORT & EXTRACT



IBM **Spectrum Discover**

- Ingests & indexes system metadata via Action Agent SDK
- Extracts labels from images
- Adds as custom tags

CURATE



Filetype: **Image**

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

Data Scientist

- Searches for images with labeled as having ‘Pothole’ feature

TRAIN

Data Scientist

- Trains a model

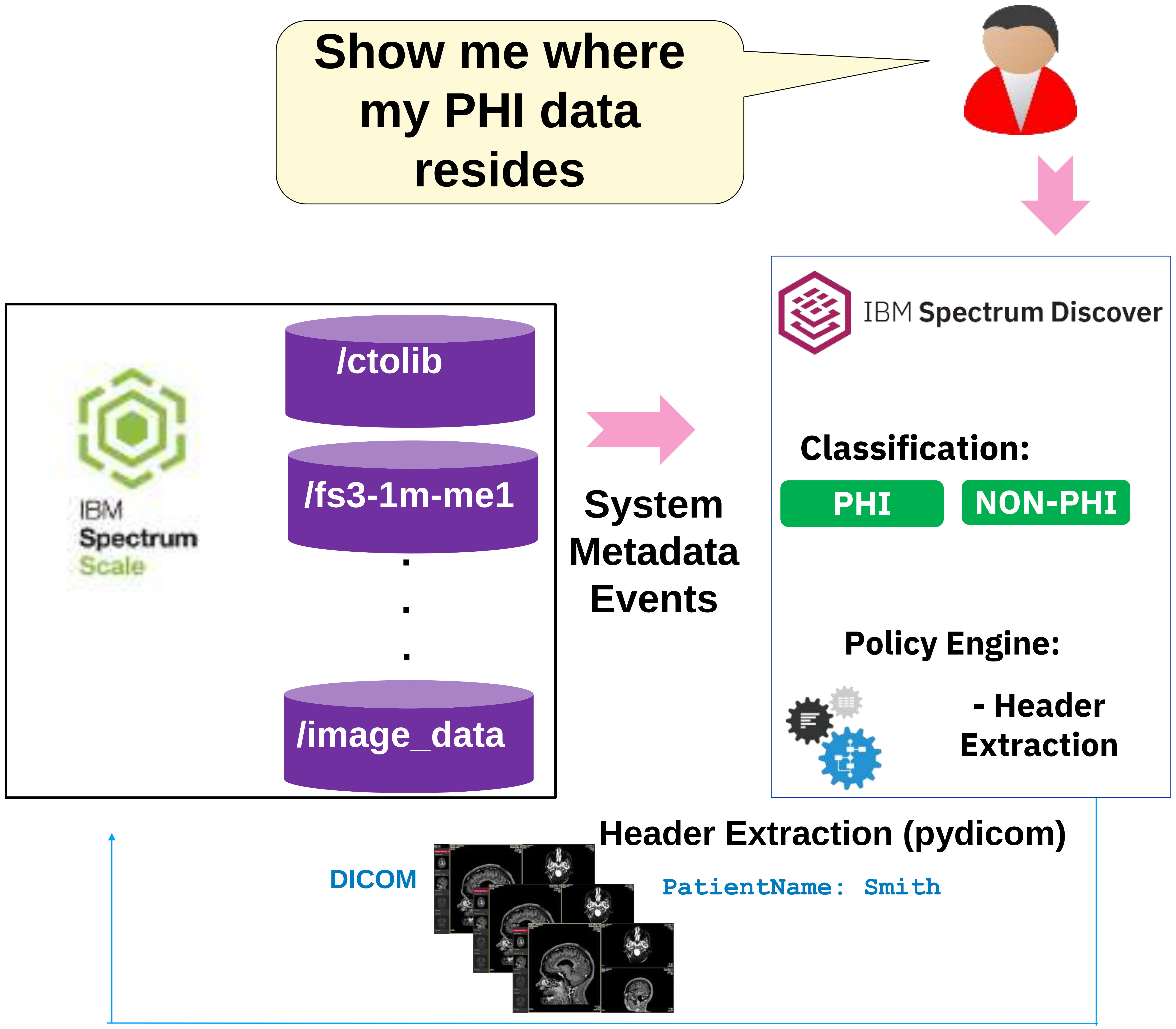
IBM **Spectrum Discover**

via Action Agent SDK

- Enriches data catalog with new tags derived from analysis

Healthcare / Life Sciences Integration

Use Case: PHI and non-PHI Classification of DICOM Images



Show me where my PHI data resides

IBM Spectrum Scale

/ctolib

/fs3-1m-me1

.
.
.

/image_data

System Metadata Events

IBM Spectrum Discover

Classification:

PHI

NON-PHI

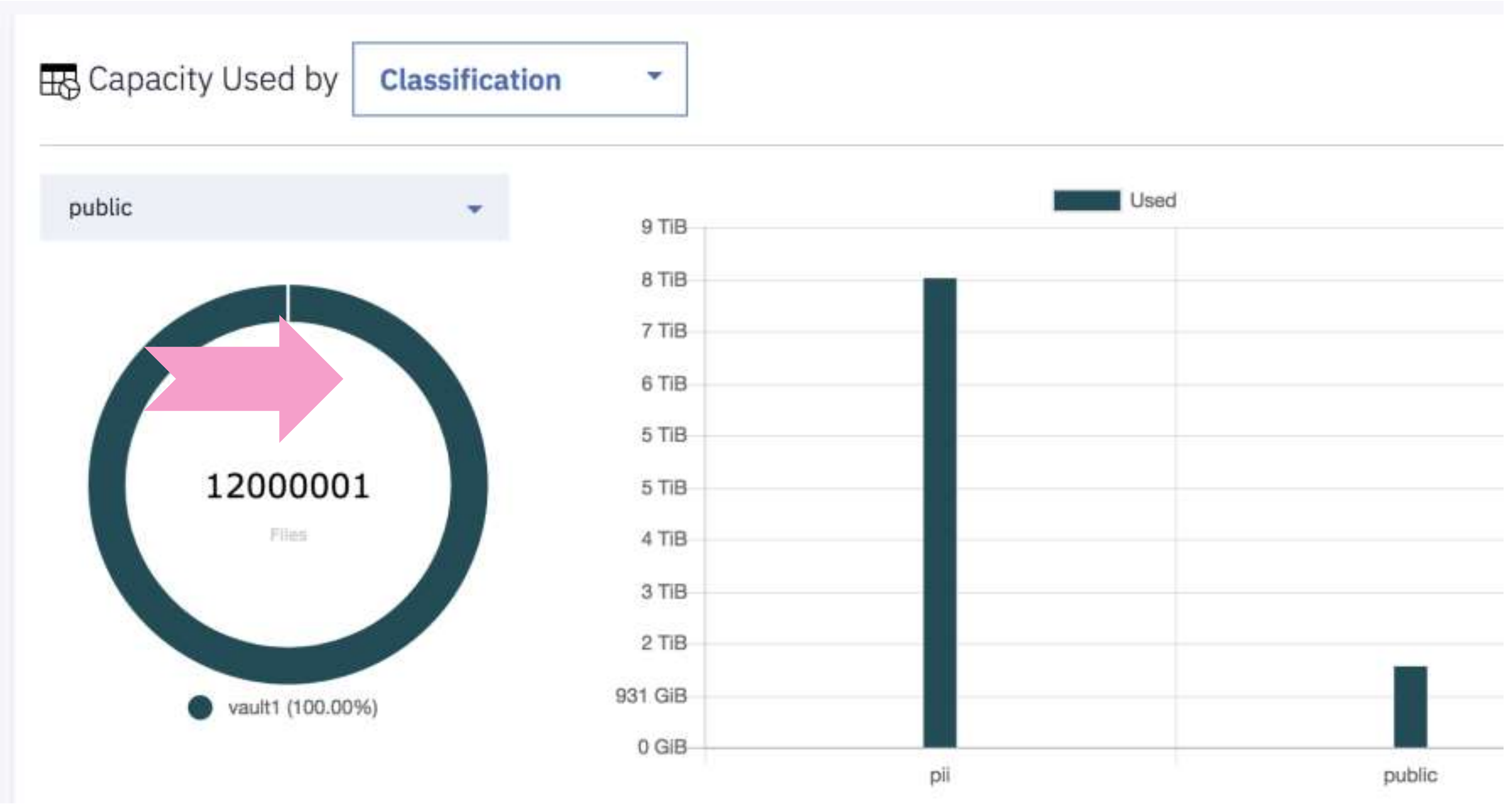
Policy Engine:

- Header Extraction

Header Extraction (pydicom)

DICOM

```
PatientName: Smith
```



Capacity Used by Classification

public

12000001 Files

● vault1 (100.00%)

IBM Spectrum Discover

Spectrum Discover Trial

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

https://www.ibm.com/us-en/marketplace/spectrum-discover

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.