



ADVANCED TECHNOLOGY GROUP (ATG)



Accelerate with ATG Webinar: IBM Fusion – Fusion Data Cataloging – Part 1: Metadata Tagging

Shu Mookerjee

ATG Senior Storage Technical Specialist

Shu.Mookerjee@ibm.com



Accelerate with ATG Technical Webinar Series

Advanced Technology Group experts cover a variety of technical topics.

Audience: Clients who have or are considering acquiring IBM Storage solutions. Business Partners and IBMers are also welcome.

To automatically receive announcements of upcoming Accelerate with IBM Storage webinars, Clients, Business Partners and IBMers are welcome to send an email request to accelerate-join@hursley.ibm.com.



2024 Upcoming Webinars – Register Here!

[Leveraging Your Data Lake with IBM Storage Fusion Data Cataloging](#)
[Part 2: April 16th, 2024](#)

[IBM GDPS 4.7 Update](#) – May 2nd, 2024

[IBM FlashSystem's FlashCore Module 4 and integrated Ransomware Threat Detection](#) - May 14th, 2024

[IBM Storage Ceph S3 Object Storage Deep Dive](#) – May 30th, 2024

Important Links to bookmark:



ATG Accelerate Site: <https://ibm.biz/BdSUFN>

ATG MediaCenter Channel: <https://ibm.biz/BdfEgQ>

Offerings

Client Technical Workshops

- **IBM DS8900F Advanced Functions:** May 8-9 in Chicago, IL
- **IBM Fusion & Ceph: A Deep Dive into Next Gen Storage:** May 15-16 in Chicago, IL
- **IBM FlashSystem Deep Dive & Advanced Functions:** May 22-23 in Atlanta, GA
- **IBM Cyber Resiliency with IBM Storage Defender:** June 5-6 in Tucson, AZ

TechZone Test Drive / Demo's

- IBM Storage Scale and Storage Scale System GUI
- IBM Storage Virtualize Test Drive
- IBM DS8900F Storage Management Test Drive
- Managing Copy Services on the DS8000 Using IBM Copy Services Manager Test Drive
- IBM DS8900F Safeguarded Copy (SGC) Test Drive
- IBM Cloud Object Storage Test Drive - (Appliance based)
- IBM Cloud Object Storage Test Drive - (VMware based)
- IBM Storage Protect Live Test Drive
- IBM Storage Ceph Test Drive - (VMware based)

Please reach out to your IBM Representative or Business Partner for more information.

***IMPORTANT* The ATG team serves clients and Business Partners in the Americas, concentrating on North America.**

Storage @ IBM TechXchange Conference 2024

(Registrations open)

October 21-24, 2024

Mandalay Bay | Las Vegas
#IBMTechXchange

Key Learnings

- Practical how-to advice
- Patterns and best practices
- Success stories, IBM PoV, proven techniques

Featured Products

IBM Storage Defender

IBM Storage Fusion

IBM Storage Scale + IBM Storage Ceph

IBM Tape + IBM SAN

IBM Storage FlashSystem + IBM Storage DS8000

Collaborate. Learn. Play.

Community

IBM Champions

User Groups

Tech Peers

Business Partners



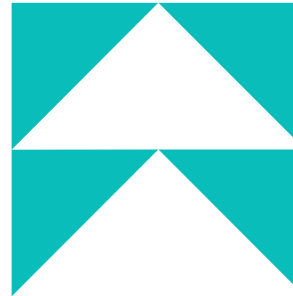
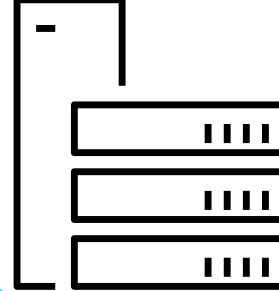
Sandbox

Network

Learn

Collaborate

Play



Accelerate your Career

Labs (Instructor-Led, Self-paced)

IBM Certification Testing

Earn up to 25 hours in CPE credits

Breakout Sessions

Trends and Directions

User Groups

Product Deep Dives

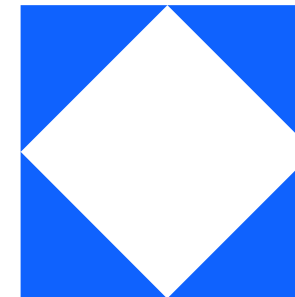
Meet the Expert

Professional Development

Show the Code

Birds of a Feather

Academic/Research



Roadmaps

Go deep with people in the know and set the stage for where IBM is going in the future



<https://www.ibm.com/community/ibm-techxchange-conference/>

Game On!



Accelerate with ATG Survey

Please take a moment to share your feedback with our team!

You can access this 6-question survey via [Menti.com](https://www.menti.com) with code 1708 6924 or

Direct link <https://www.menti.com/alwhyze7z1gz>

Or

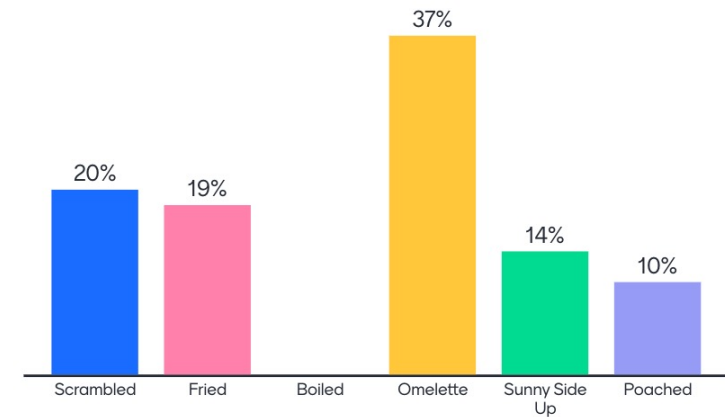
QR Code



Join at [menti.com](https://www.menti.com) | use code 1708 6924

 Mentimeter

What's the best way to eat eggs?



ADVANCED TECHNOLOGY GROUP (ATG)



Accelerate with ATG Webinar: IBM Fusion – Fusion Data Cataloging – Part 1: Metadata Tagging

Shu Mookerjee

ATG Senior Storage Technical Specialist

Shu.Mookerjee@ibm.com

Meet the Speaker and Panelists



Shu Mookerjee is a Level 2 Certified Technical Specialist with over twenty years at IBM, working in a variety of roles including sales, management and technology. For the last decade, he has focused exclusively on storage and has been the co-author of four (4) Redbooks. Currently, Shu is part of the Advanced Technology Group where he provides education, technical guidance, Proofs of Concept and Proofs of Technology to IBMers, business partners and clients.



Norm Bogard is a Senior IT Specialist in the Public Sector for IBM Storage Software. He was over 30 years of storage industry experience specializing in hybrid cloud software, file access, and metadata management



Lloyd Dean is an IBM Principal Storage Technical Specialist in IBM Storage Solutions. Lloyd has held numerous senior technical roles at IBM during his 22 plus years at IBM. Lloyd most recently is leading efforts in the Advanced Technology Group as the IBM Storage for Red Hat OpenShift focal and as a Hybrid Cloud storage solution SME covering IBM Block, File and Object storage solutions and their use cases supporting IBM Cloud Paks.

Agenda

- Goals and Objectives
- IBM Fusion Data Services
- Fusion Data Cataloging
- Metadata Tagging and Policies
- Other Cool Stuff (Miscellaneous)
- Demo

Goals and Objectives

Objective:

Introduce Fusion Data Cataloging and the metadata tagging concept

We WILL:

- Provide an overview of the Fusion Data Cataloging service
- Review its architecture at a high level
- Walk through simple set up and usage

We WILL Not:

- Do a deep dive into various metadata tagging policies
- Spend time on complete deployment
- Talk about competitive products or solutions

Note: This is part of a Two-Part Series.

- **Part One will cover Tagging and Sorting**
- Part Two will cover Tagging and Moving

Fusion Data Services

IBM Fusion provides data services for containerized/OpenShift applications

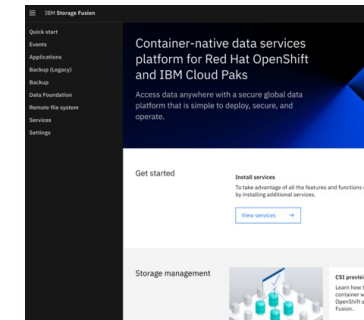
IBM Fusion HCI



NextGen Data Services

- Data Persistence
- Data Security
- Data Discovery
- Data Mobility
- Data Resilience

IBM Fusion Software



Fusion Data Services – Data Discovery

Data Persistence



Keeps data when the power goes off or the host goes down

Data Security



Protecting data from cyber attacks and enabling a quick recovery

Data Discovery



Understanding, organizing and cataloging data

Data Mobility



Enabling data to be put in the right place with the right policies and controls

Data Resilience



Copying and backing up data in order to keep applications running

Fusion Data Cataloging – Why Catalog Data?

- More than 80% of an organization’s data is in the form of “Unstructured Data”
- So what is unstructured data? Depends on who you ask!
 - Data with no a pre-defined data model or structure
 - Data stored in a structural database format
 - Data fit for human consumption (.pdfs, .gifs, .wav, etc...)
- Has inherent management challenges. It’s difficult to:
 - Pinpoint & activate relevant data for large-scale analytics
 - Map data to business, projects
 - Remove redundant, trivial & obsolete data
 - Identify & classify sensitive data
- Data “prep time” still dominates analyst and data scientist’s time

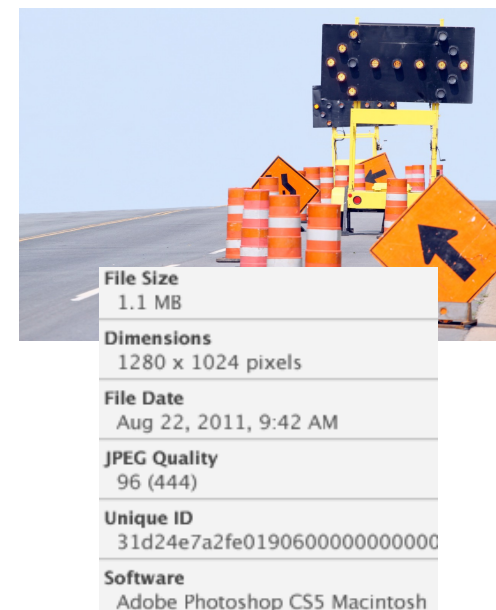


Fusion Data Cataloging – Unstructured Data and Metadata

- A data object* may be unstructured...but its metadata isn't!
- Metadata is structured information about the data object (“the data about the data”)
- 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

*Note: "data object" = "the data with the metadata", not necessarily an S3 object

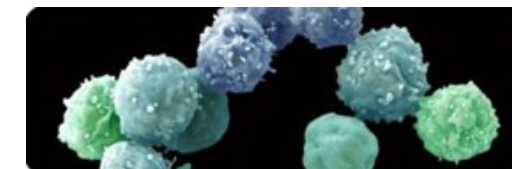
Images



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, antibody, Organism,

Natural Language Processing



Fusion Data Cataloging - Overview

File, Object, Backup, and Archive Storage



IBM
Spectrum
Scale



IBM
Spectrum
Protect



IBM Cloud
Object
Storage



IBM
Spectrum
Archive



S3



NetApp™



ceph

Data Insight



IBM Spectrum Discover



Search



Reporting



Dashboard

- Simple to deploy (bare metal, vmWare, **OpenShift**)
- Metadata curation
- Custom metadata tagging
- Automatic indexing
- Policy-Engine
- Action Agent API

Activation & Optimization

Large-Scale Analytics

- Data discovery
- Dataset identification
- Data pipeline progression

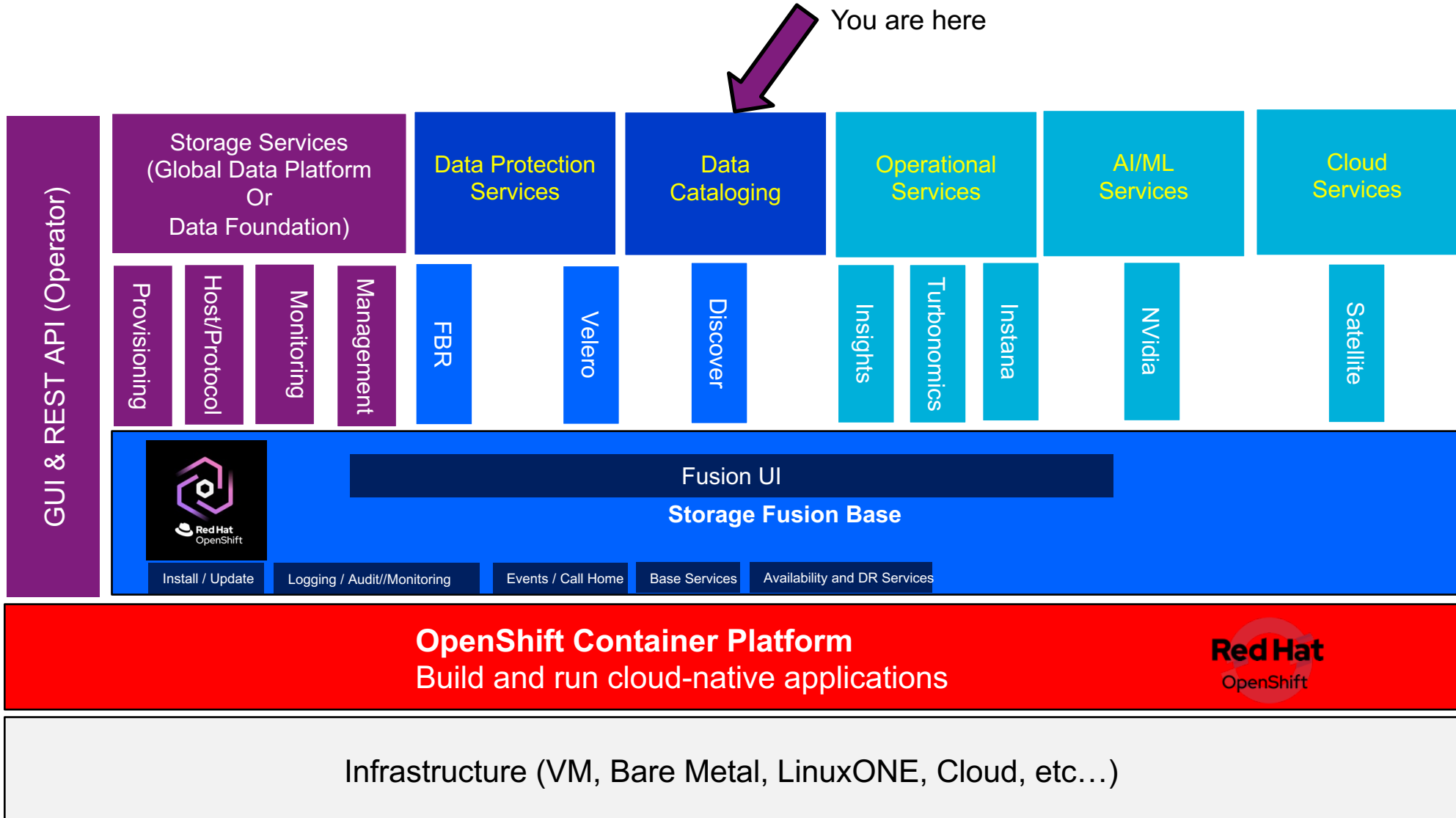
Data Governance and privacy

- Data inspection
- Data classification
- Data clean-up

Data Optimization

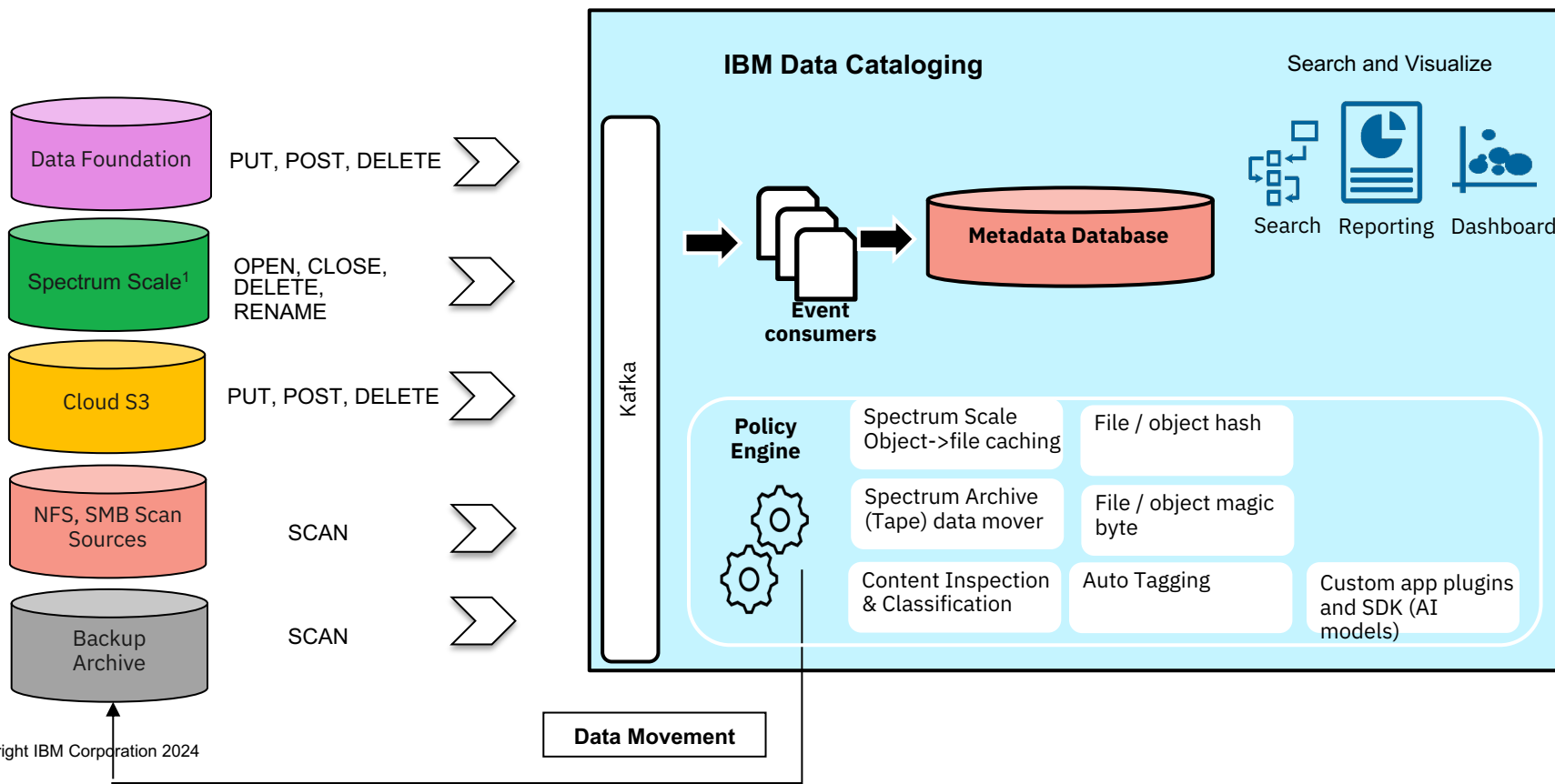
- Archive / tiering
- Duplicate data removal
- Trivial data removal

Fusion Data Cataloging - Architecture



Fusion Data Cataloging – Process and Architecture

- Metadata scanned and processed through Apache Kafka in real time
- Data is run through policy engine (as appropriate)
- Data is tagged, catalogued and sorted via APIs to the policy engine service
- Store pointers in the metadata database (DB2 Data Warehouse)
- Curated and ready for post scan analytics, reporting, etc...

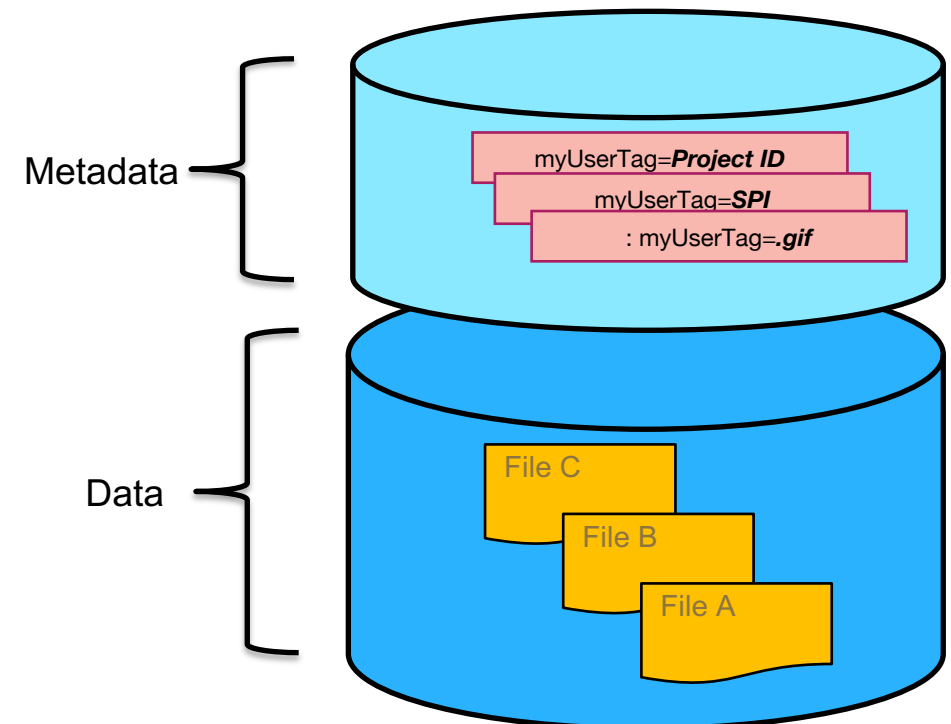


Currently Supported Data Sources

- IBM Storage Scale
- IBM Storage Scale System
- IBM Cloud Object Storage
- IBM Storage Protect
- IBM Storage Archive
- IBM Ceph Storage
- NetApp Storage Solutions
- Dell EMC Isilon Scale Out Network Attached Storage
- Amazon Simple Storage Service (Amazon S3)
- NFSv3 and NFSv4

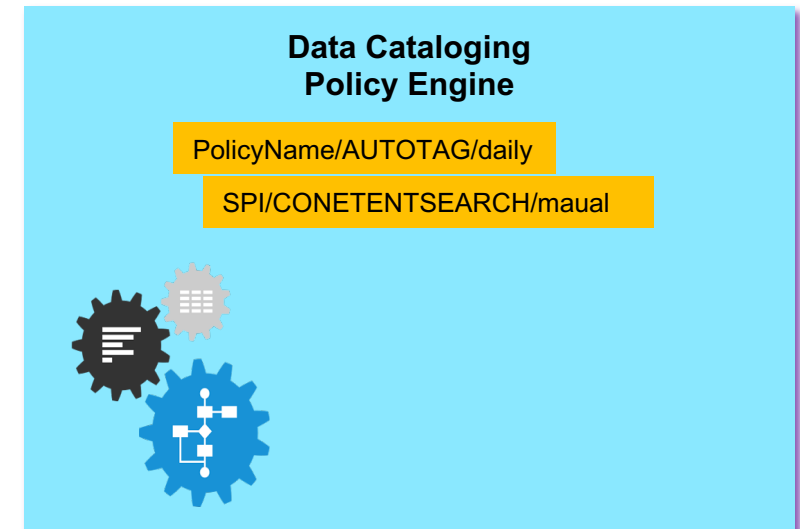
Metadata Tagging

- Data Cataloging can enrich the metadata from supported platforms with more information by using custom tags, policies, action agents
- A tag is a custom metadata field (or key:value pair) that is used to supplement storage system metadata with organization-specific information
- Two types of available tags:
 - **Categorization Tag:** Contain values such as project, department, and security classification.
 - Can be open or restricted.
 - If it is open, listed selections can be used.
 - If it is restricted, selection is limited to true or false.
 - **Characteristic Tag:** Can contain any value that is needed to describe or classify the object.
 - Can contain long descriptive values
 - Size limit is 4 KB



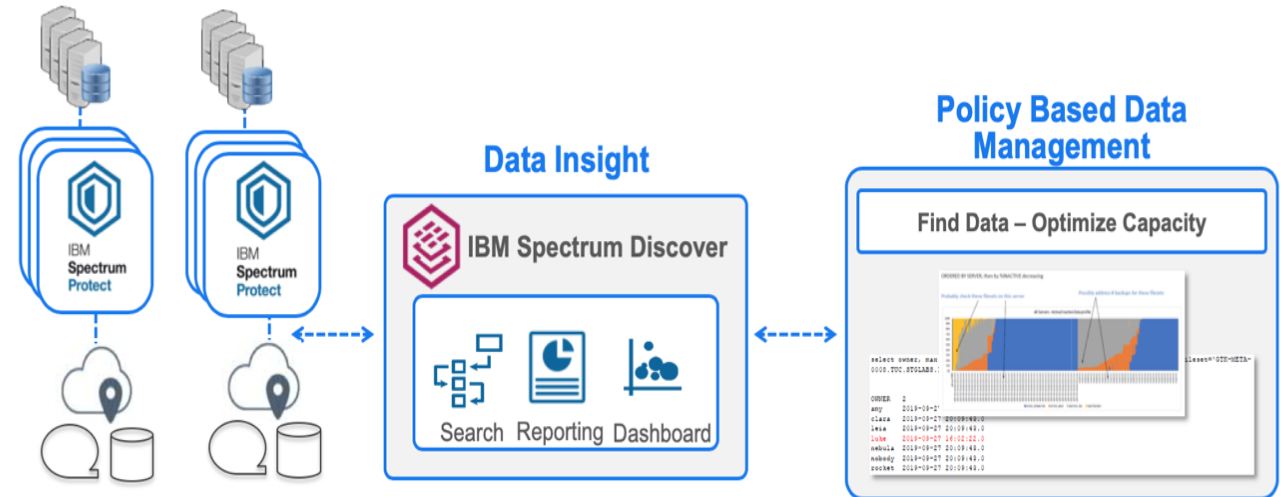
Metadata Tagging - Policies

- Policies offer a way to perform specific actions on a filtered set of records. These can be immediate or scheduled
- The policy management is a RESTful API service designed to create, list, update, and delete policies
- There are three types of policies supported:
 - **AUTOTAG**
 - Tags records based on filter criteria with a pre-defined set of tags
 - Logical representation would be: if (<filter>) then <tag> = <value>
 - **CONTENT SEARCH**
 - Uses the built-in content inspection capabilities to extract content from metadata and index it automatically
 - Leverages existing or custom RegEx expressions
 - **DEEP-INSPECT**
 - Engages external agents to provide specific metadata tagging and inspection
 - Agents can be found in the software developer kit (SDK) available at [GitHub](#). A sample app is available on [Dockerhub](#)



Other Cool FDC Stuff – Backup Solutions

- Integration with Storage Protect
- Insight into existing backups
- Quickly find and activate cold data in backup/archive for analytics and AI
- Automatically identify duplicate files for better storage utilization



Other Cool FDC Stuff – Sizing

- Sizing is dependent on the ***number of data objects to be ingested (workload), not capacity***
- Key driver of FDC sizing is Db2 Data Warehouse
- Deployment available in “T-shirt” size type configurations (S, M, L)
- Sizing can scale linearly

Deployment	Node	vCPU	RAM	Disk Space	Network	Storage	Workload
All Instances	Two Control Nodes	2 x 8	2 x 16 GB	100 GB	10 GB		
Starter Profile	Two Worker Nodes	2 x 16	2 x 32 GB	120 GB	10 GB	500 GB	50 M
Mid Range Profile	Two Worker Nodes	2 x 34	2 x 64 GB	120 GB	10 GB	2.5 TB	1 B
Large Profile	Two Worker Nodes	2 x 380	2 x 814	120 GB	10 GB	21.4 TB	12 B

Other Cool FDC Stuff – Some Final Thoughts...


Though Fusion Data Catalog can provide a variety of functions, it does not/cannot:

- EVER scan or open the **actual data object**
 - All the information it ingests comes from the metadata
 - The data object itself remains untouched
- Copy the metadata to the Db2 database
 - It does, however, store pointers
 - Thus the quantity of data objects is more important than their consumed capacity
- Move or manipulate the data object
 - It relies on external “data mover” applications
 - Leverages APIs to connect with those applications
- Provide performance data (see below)



Remember...if it ain't in the metadata, FDC can't tag it, sort it or provide information on it!

Shameless Plug! In-Person Fusion and Ceph Workshop – Chicago, IL



ADVANCED TECHNOLOGY GROUP (ATG)

IBM Fusion and Ceph ATG Workshop: A Deep Dive into Next Gen Storage May 15-16, 2024 – Chicago, IL

More organizations are moving towards application modernization, hybrid cloud infrastructure and AI to meet the rapidly increasing speed of business. A key component of this transformation is the deployment of lightweight containerized applications. However, these applications require data services that are consistent, persistent, and highly functional. IBM Storage Fusion is designed specifically to serve those needs.

Join us for a two-day in-person technical workshop led by Subject Matter Experts from the Advanced Technology Group (ATG).

Description: This workshop will cover a range of technical topics including:

- > IBM Storage Fusion concepts and architecture
- > IBM Storage Ceph concepts and Fusion Data Foundation integration
- > Implementation and installation options for both Fusion HCI and Fusion Software
- > Data protection, backup and [restore](#)
- > Data cataloging and metadata tagging
- > Using an External Ceph Storage Cluster with Fusion Data Foundation
- > Deployment use cases for IBM Watsonx, AI and Machine Learning

In addition, there will be live demos and hands-on labs.

Audience: Clients along with their respective Business Partner and IBM Host

Dates: Wednesday, May 15th & Thursday, May 16th, 2024



Times: [May 15](#): 9:00am – 4:00pm and [May 16](#): 9:00am – 3:00pm

Location: IBM Chicago, 71 S. Wacker Drive, 6th Floor, Chicago IL 60606

Speakers: Shu Mookerjee, Senior Storage Technical Specialist, Lloyd Dean, Principal Brand Technical Specialist and John Shubeck, Senior Storage Technical Specialist

Important Notes:

- > This is an IBM nomination event. Please check with your IBM or BP representative on how you may qualify to attend.
- > Customers will be responsible for their own travel and lodging arrangements and expenses.



ADVANCED TECHNOLOGY GROUP (ATG)

IBM Fusion and Ceph ATG Workshop: A Deep Dive into Next Gen Storage Agenda

Wednesday, May 15 – Fusion and Ceph Principles and Concepts

- Intro to Fusion - Part 1 - Containers, OpenShift & App Modernization
- DEMO - OpenShift Walkthrough
- Intro to Fusion - Part 2 - Concepts and Services
- DEMO - Fusion Walkthrough
- Break
- Deploying Ceph Part 1 (Introduction to Ceph, architecture, access protocols)
- Lunch
- Deploying Ceph Part 2 (S3 object storage)
- Lab - Implementing S3 object storage with IBM Storage Ceph
- Break
- Application Modernization Support with Fusion
- DEMO - OpenShift Virtualization
- Lab - Deploying Fusion with External Ceph

Thursday, May 16 – Data Services and Use Cases

- Data Protection Services Overview
- Backup and Restore
- DEMO - Fusion Backup and Restore
- Data Protection and Resilience in Ceph
- Lunch
- Fusion Replication & AFM
- Fusion Data Cataloging
- DEMO - Data Cataloging and ILM
- Watson and Ceph
- Fusion HCI and watsonx.data
- Close

Alright Stop.



Demo Time!

Accelerate with ATG Survey

Please take a moment to share your feedback with our team!

You can access this 6-question survey via [Menti.com](https://www.menti.com) with code 1708 6924 or

Direct link <https://www.menti.com/alwhyze7z1gz>

Or

QR Code

