ADVANCED TECHNOLOGY GROUP (ATG)

Accelerate with ATG Webinar: IBM Fusion – Fusion Data Cataloging – Part 2: Data Moving

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
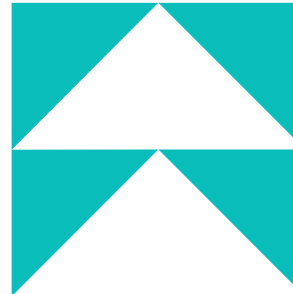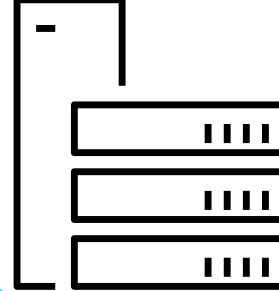
IBM Champions

User Groups

Tech Peers

Business Partners

Sandbox

Network   Learn

Collaborate   Play

Roadmaps

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

Breakout Sessions

Trends and Directions   User Groups

Product Deep Dives   Meet the Expert

Professional Development   Show the Code

Birds of a Feather

Academic/Research

Accelerate your Career

Labs (Instructor-Led, Self-paced)

IBM Certification Testing

Earn up to 25 hours in CPE credits

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 with code 1708 6924 or

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

Or

QR Code

FUN Question: What is the best type of donut?
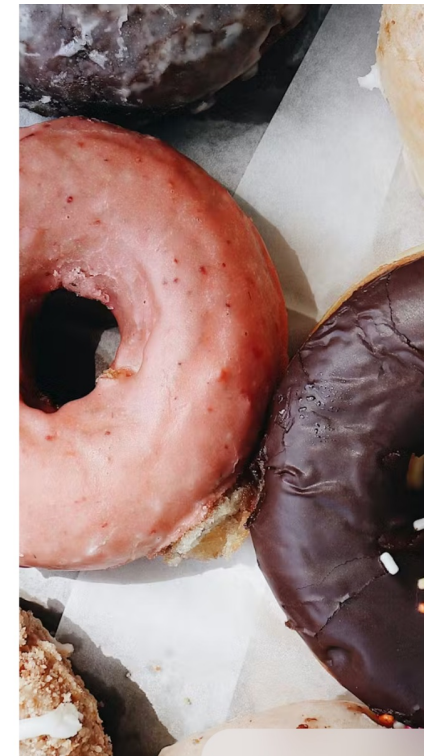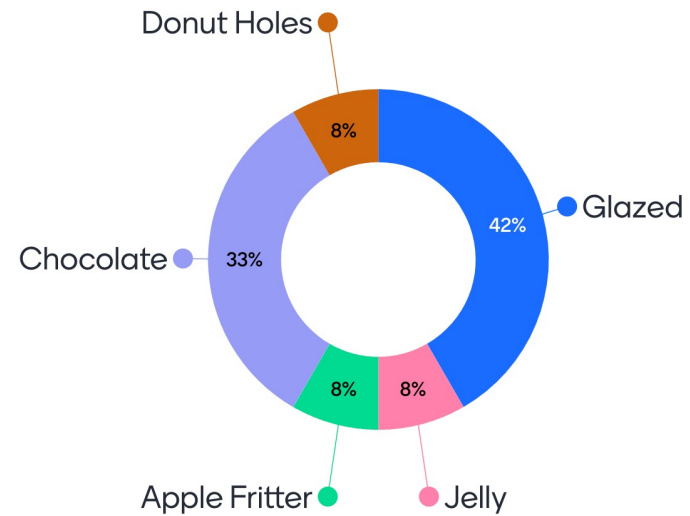
- Donut Holes — 8%
- Glazed — 42%
- Chocolate — 33%
- Jelly — 8%
- Apple Fritter — 8%

ADVANCED TECHNOLOGY GROUP (ATG)

Accelerate with ATG Webinar: IBM Fusion – Fusion Data Cataloging – Part 2: Data Moving

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.

ADVANCED TECHNOLOGY GROUP (ATG)

Agenda

- Goals and Objectives

- Intro to IBM Fusion Data Services

- A Recap Fusion Data Cataloging

- Fusion Data Cataloging and Data Placement Technology

- Demo

© Copyright IBM Corporation 2024

Goals and Objectives

Objective:
Demonstrate how to leverage Fusion Data Cataloging for data placement, tiering and migration

Two Big Assumptions:
- You either attended or watched a replay of last week's session ("IBM Fusion – Fusion Data Cataloging – Part 1")
- You have some basic understanding of IBM Fusion

We WILL Not:
- Do a deep dive of Fusion or Fusion Data Cataloging (though we WILL do a recap)
- Talk about competitive products or solutions
- Cover setup, installation or implementation

Note: This is part of a Two-Part Series.
- Part One will cover Tagging and Sorting
- Part Two will cover Tagging and Data Placement

IBM Fusion Data Services

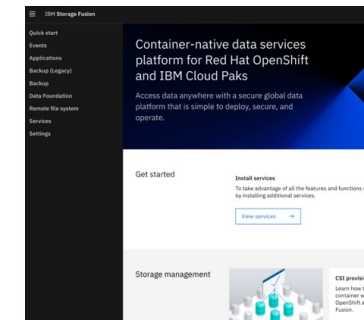*IBM Fusion provides data services for containerized/OpenShift applications*

IBM Fusion HCI

IBM Fusion Software

NextGen Data Services

Data Persistence

Data Security

Data Discovery

Data Mobility

Data Resilience

IBM Fusion Data Services

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

A Recap of Fusion Data Cataloging – What Is It?

Fusion Data Cataloging is:

- Built on the same technology as IBM Spectrum Discover

- Available as a data service for IBM Fusion

- Designed to provide management and understanding of unstructured data
  - Data with no pre-defined tabulated model or structure
  - Data stored in a structural database format
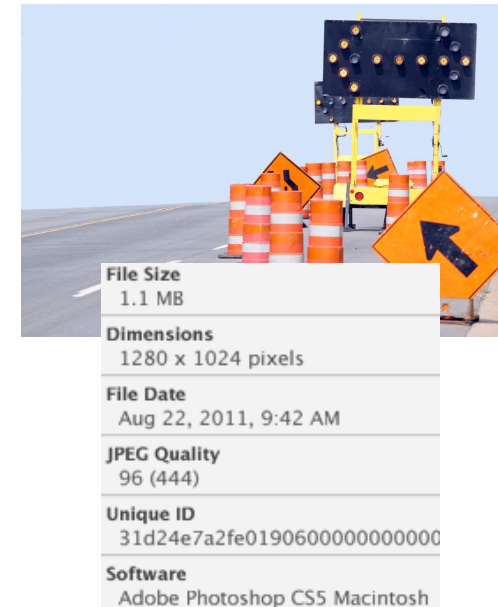 - Data fit for human consumption (.pdfs, .gifs, .wav, etc…)

A Recap of Fusion Data Cataloging – What's It For?

- Fusion Data Cataloging enables users to make sense of their unstructured data by inspecting, tagging and sorting the 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")

- <u>Who</u>, <u>what</u>, <u>when</u>, <u>where</u>, and <u>why</u> of the data object (account, container, object, stream, dir, file, etc…)

- 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

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

System Metadata

Location
Size
Owner
Group
Permissions
Last-Modified
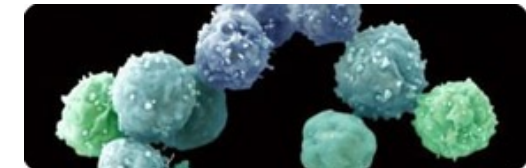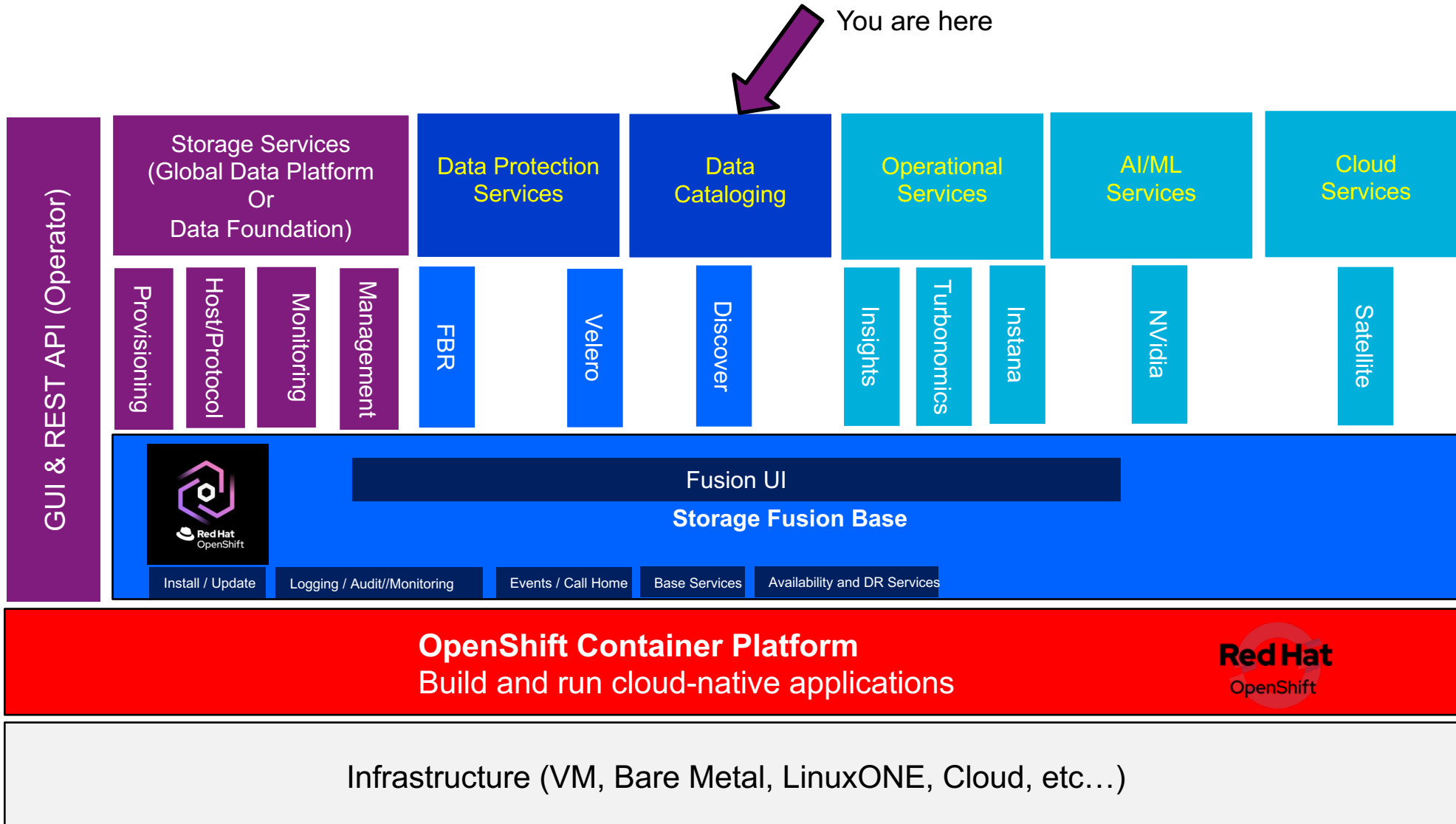...

Biomedical

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

Natural Language Processing

A Recap of Fusion Data Cataloging – Where Does it Live?

You are here

GUI & REST API (Operator)	Storage Services (Global Data Platform Or Data Foundation)				Data Protection Services		Data Cataloging	Operational Services			AI/ML Services	Cloud Services
	Provisioning	Host/Protocol	Monitoring	Management	FBR	Velero	Discover	Insights	Turbonomics	Instana	NVidia	Satellite

Fusion UI

Storage Fusion Base

Install / Update   Logging / Audit//Monitoring   Events / Call Home   Base Services   Availability and DR Services

OpenShift Container Platform
Build and run cloud-native applications

Red Hat OpenShift

Infrastructure (VM, Bare Metal, LinuxONE, Cloud, etc…)

A Recap of Fusion Data Cataloging – How Does it Work?

- 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…
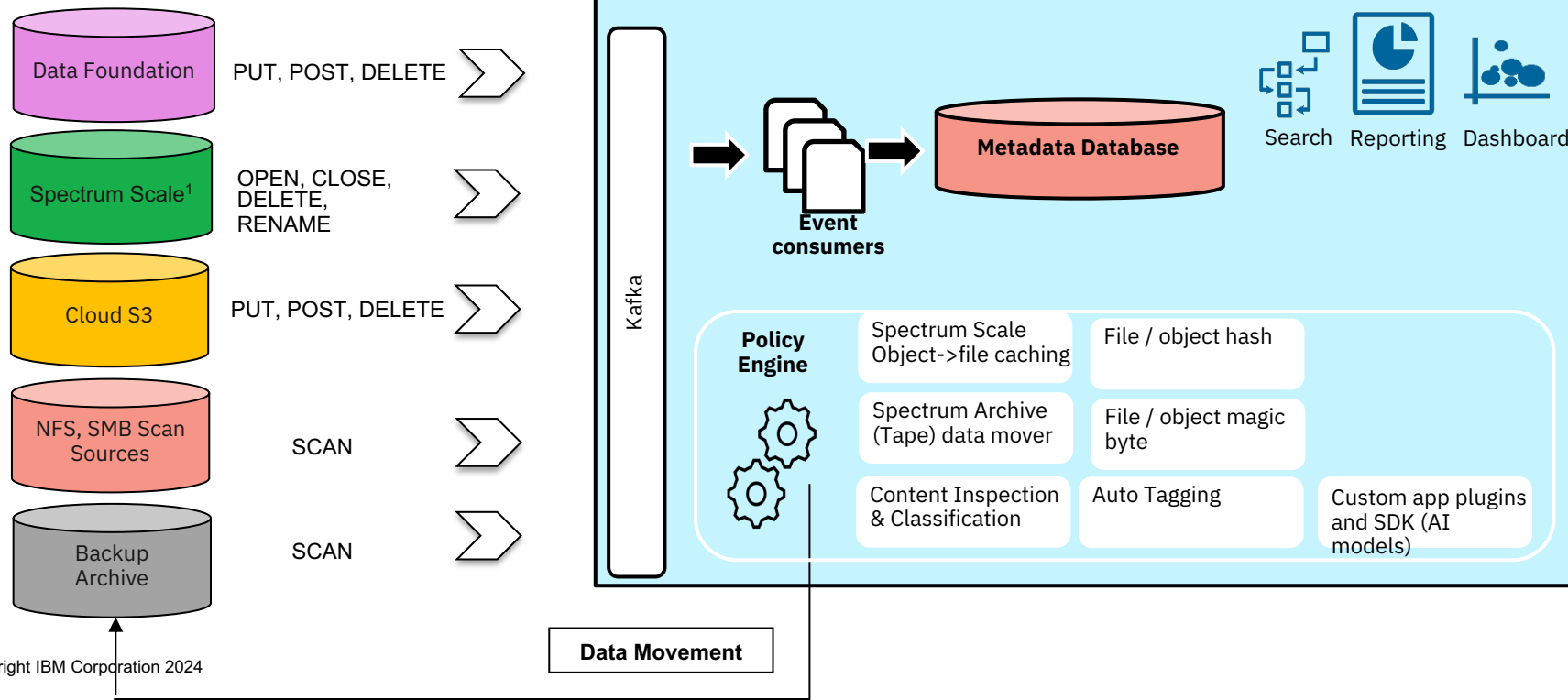
**Currently Supported Data Sources**

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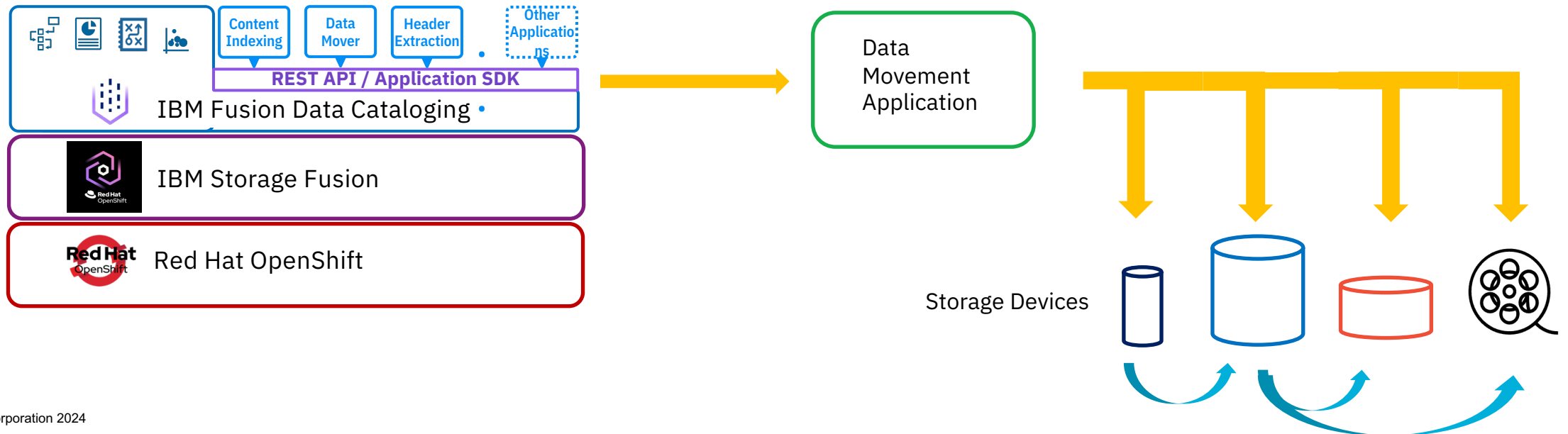- Amazon Simple Storage Service (Amazon S3)
- NFSv3 and NFSv4

IBM Data Cataloging

Search and Visualize — Search, Reporting, Dashboard

Kafka → Event consumers → **Metadata Database**

Policy Engine: Spectrum Scale Object->file caching, File / object hash, Spectrum Archive (Tape) data mover, File / object magic byte, Content Inspection & Classification, Auto Tagging, Custom app plugins and SDK (AI models)

Data Foundation — PUT, POST, DELETE
Spectrum Scale[1] — OPEN, CLOSE, DELETE, RENAME
Cloud S3 — PUT, POST, DELETE
NFS, SMB Scan Sources — SCAN
Backup Archive — SCAN

Data Movement

FDC and Data Placement Technology – So What? Who Cares?

- Actively enables data optimization!

- Tagging metadata enables data to be sorted and scrutinized by use case instead of technology

- Links Policy Engine to the data migration applications that enable placement directly from FDC interface
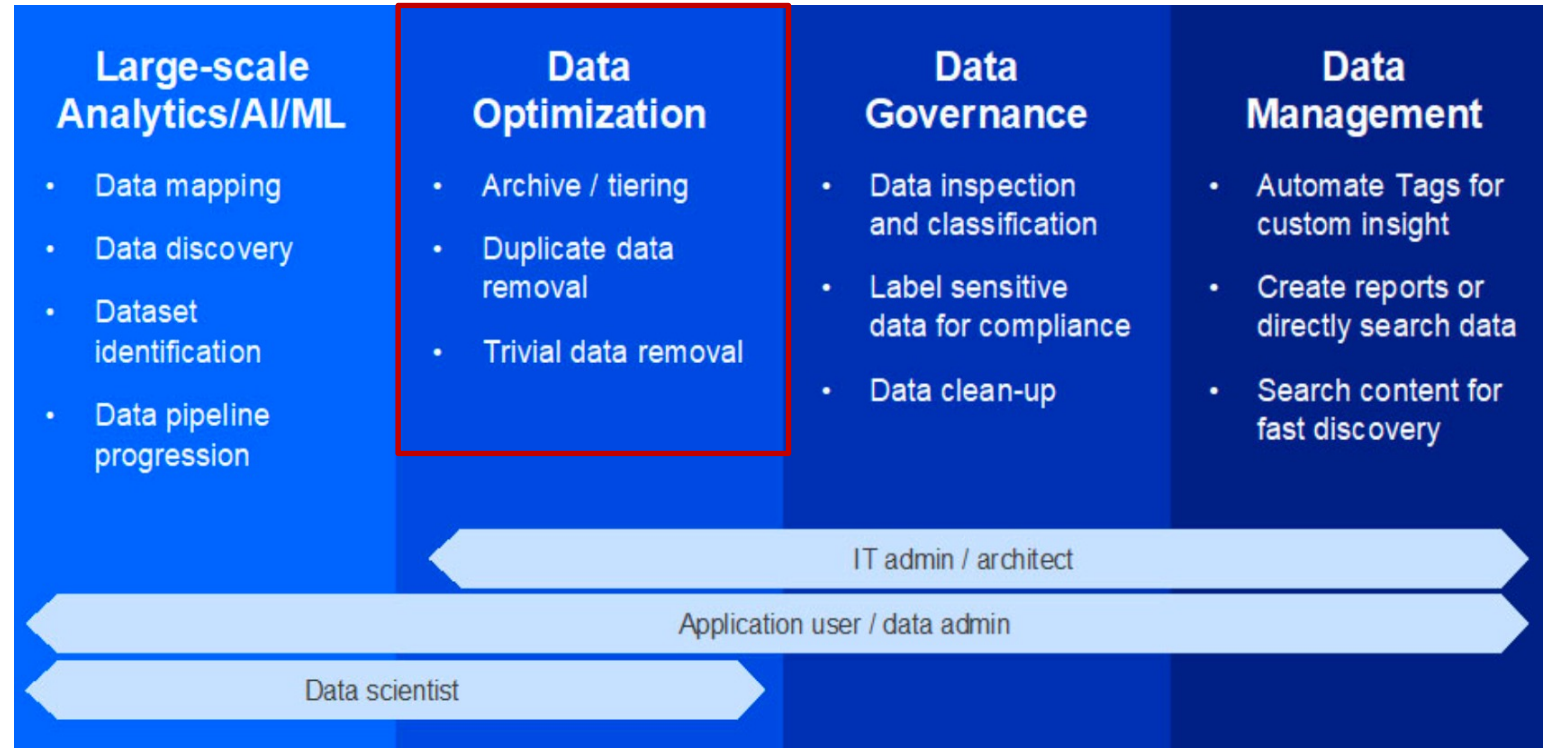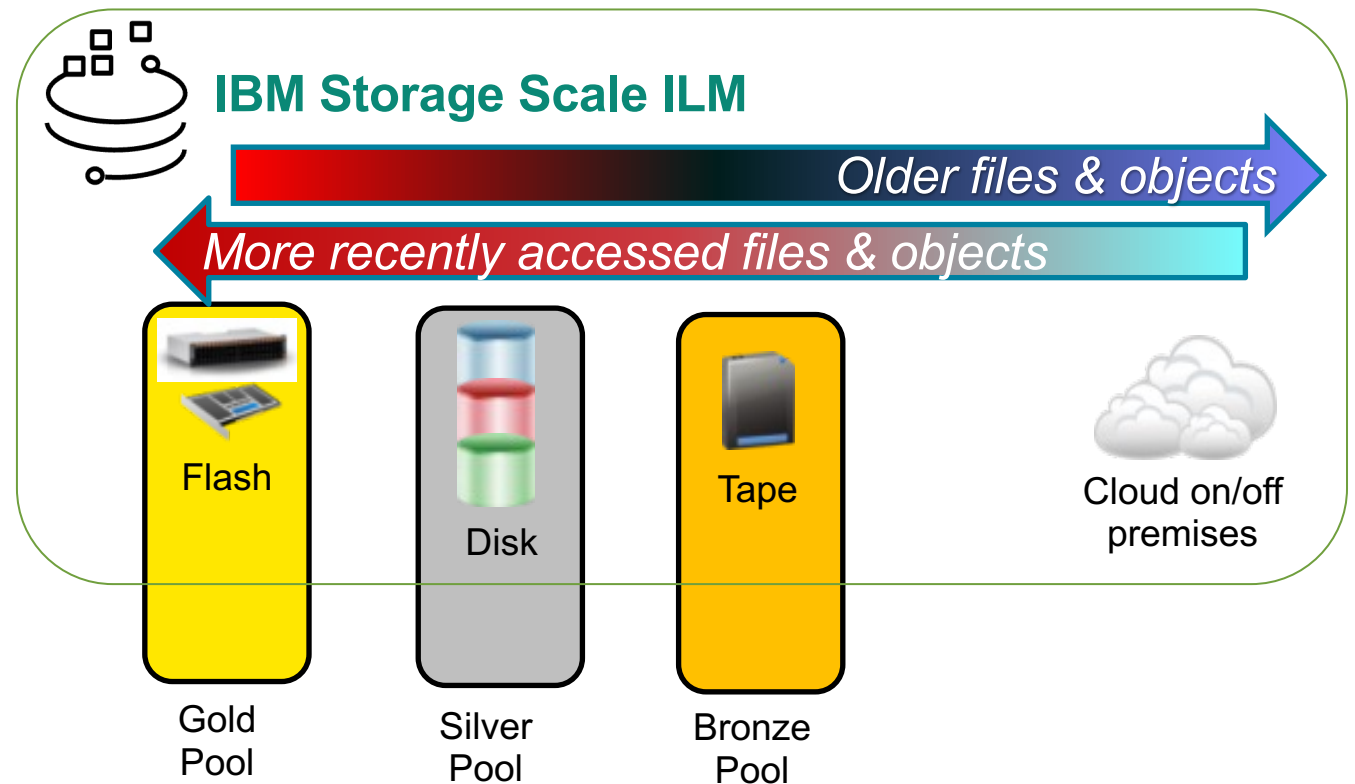
FDC and Data Placement Technology - Architecture

- Fusion Data Cataloging leverages extensible APIs to communicate with data placement applications. (**NOTE: FDC itself does not move the data)**

- Currently, only two are supported:
  - IBM Storage Scale Information Lifecycle Management (ILM)
  - Moonwalk

- Data placement application will communicate with the storage targets via their connection process

- In all cases data placement policies are created and managed through FDC while the data movement application will perform that actual operation

Content Indexing · Data Mover · Header Extraction · Other Applications

REST API / Application SDK

IBM Fusion Data Cataloging

IBM Storage Fusion

Red Hat OpenShift

Data Movement Application

Storage Devices

FDC and Data Placement Technology – Scale Information Lifecycle Management

- IBM Storage Scale Information Lifecycle Management ("Scale ILM") enables tiering and data placement within a Storage Scale cluster

- Placement "pools" can be created in Scale based on file, fileset or media (see example)

- Placement policies can be created at the file or fileset level to tier data between pools based on attributes such as owner, group, size, pathname, fileset, last access/modified time, "file heat", extended attributes, etc

- Fusion Data Cataloging provides the advanced data tiering function through the IBM Storage Scale connection via metadata tagging

IBM Storage Scale ILM

Older files & objects

More recently accessed files & objects

Flash

Disk

Tape

Cloud on/off premises

Gold Pool

Silver Pool

Bronze Pool

FDC and Data Placement Technology – Moonwalk

- Moonwalk's "Dynamic Workload Placement" solution is a large-scale heterogeneous data management system

- Supports multiple unstructured data storage sources such as IBM, NetApp, Dell and AWS

- Automates and manages the movement of data from primary storage to secondary storage or archival targets such as object stores, tape, or cloud storage services

- Integrates seamlessly with Fusion Data Catalog to provide data mobility across devices based on metadata policy

FDC and Data Placement Technology – Demo Environment

1. Fusion Data Cataloging (FDC) locates files within the Scale cluster based on filter used ("Newton")
2. ILM Pools have already been created on Scale ("system", "archivestaging", "ETL")

    Note: The "archivestaging" node ***must be*** used as "Host" in FDC Data Connection
3. FDC passes data management instructions via Scale management policy
4. Scale initiates data movement of files listed in policy passed by FDC
5. FDC catalog is updated after 30 minutes of the completion of data management policy carried out by Scale.
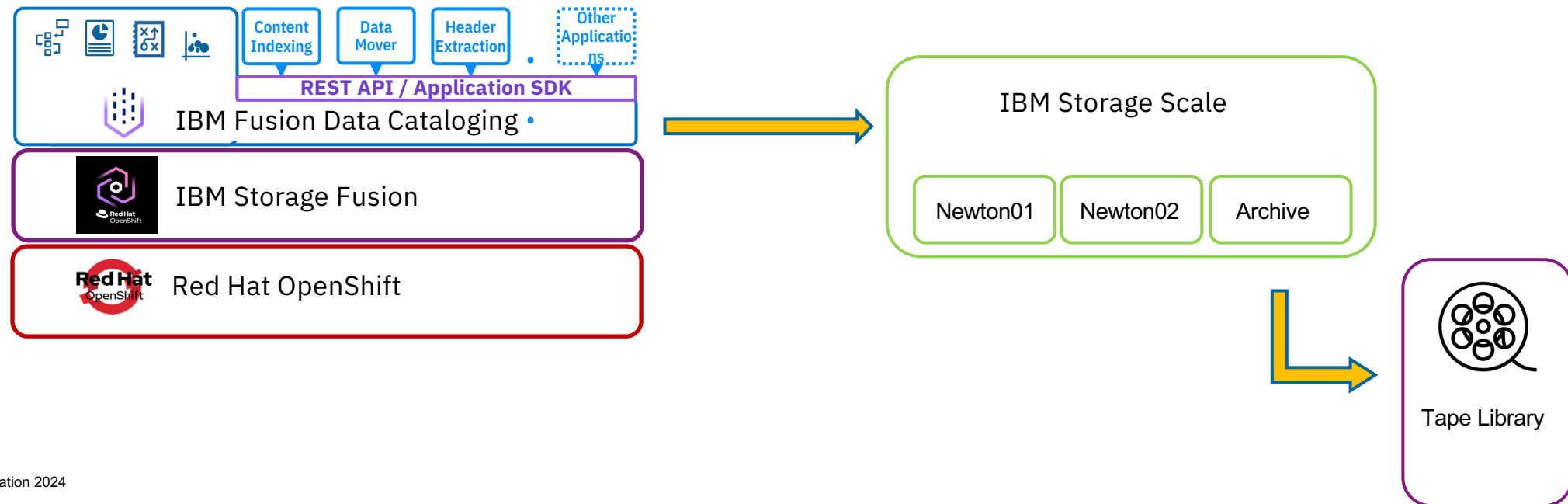
Shameless Plug! In-Person Fusion and Ceph Workshop – Chicago, IL

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 with code 1708 6924 or

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

Or

QR Code