Best practices for data collection - a critical foundation for obtaining valuable data insights

Jason Mathew
Offering Manager, Watson Knowledge Catalog

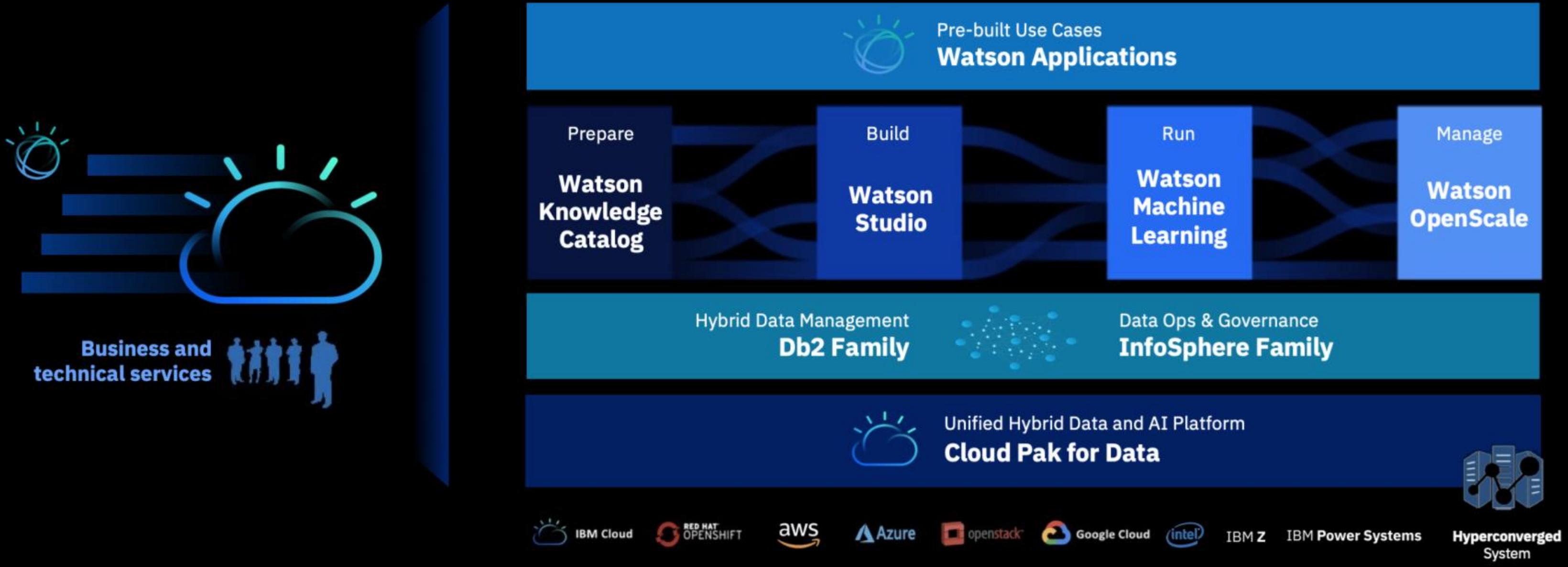
IBM Data and AI

David Wohlford
Product Marketing Manager, Spectrum Discover

IBM Systems

Metadata unlocks data by making it visible and understandable.

Watson™ Knowledge Catalog provides metadata management for the IBM Data and AI Portfolio



Business and technical services

Pre-built Use Cases
Watson Applications

Prepare
Watson Knowledge Catalog

Build
Watson Studio

Run
Watson Machine Learning

Manage
Watson OpenScale

Hybrid Data Management
Db2 Family

Data Ops & Governance
InfoSphere Family

Unified Hybrid Data and AI Platform
Cloud Pak for Data

Benefits of metadata management

Regulatory compliance

Metadata management conducted on a unified platform that provides stewardship, data lineage, and impact analysis services is the best assurance that an organization can validate and demonstrate that the data reported is true.

Productivity and discovery

Data is abundant. Much of it comes from existing systems and data stores for which no documentation exists, or the documentation does not reflect the changes and updates of those systems and data stores.

Mitigating risk

Metadata management provides the measure of trust that businesses need. Through data lineage and impact analysis, businesses can know the accuracy, completeness and currency of the data used in their planning or decision-making models.

DataOps Impact

Data Inventory Example

85%

Reduction in business glossary creation time

90%

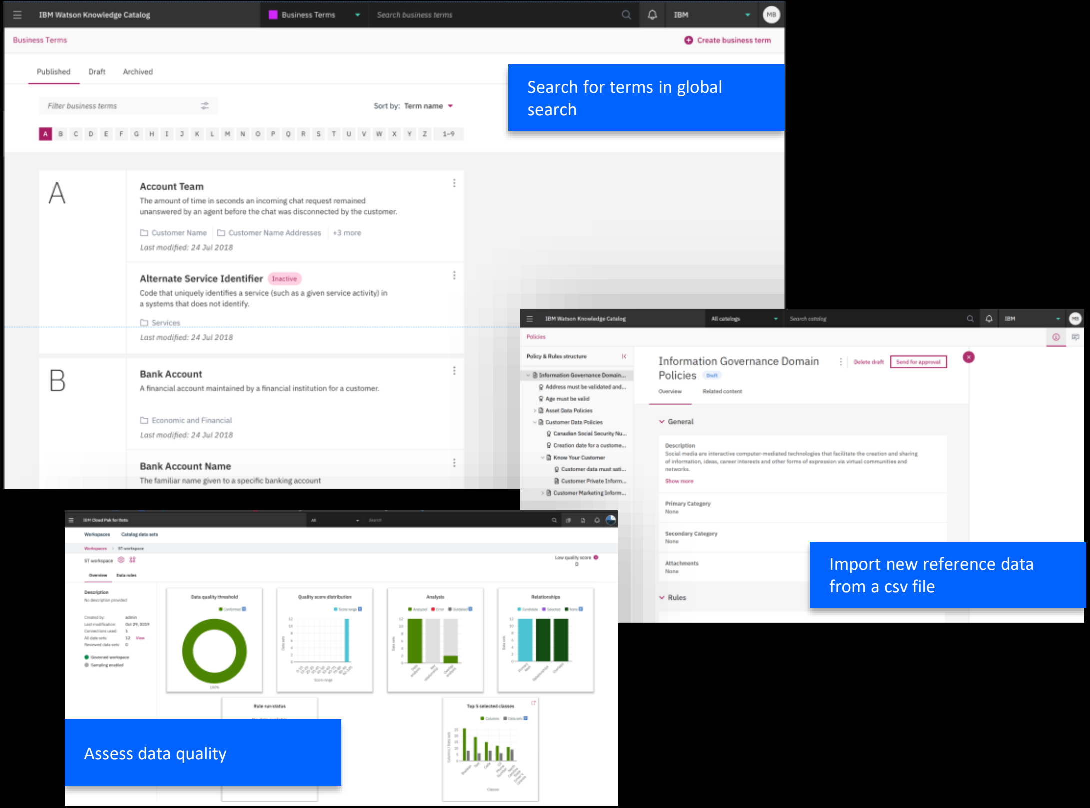
Reduction in time to discover metadata and assign terms

200,000

Number of technical assets across multiple clouds discovered in less than 5 mins

Watson™ Knowledge Catalog

Open and intelligent data catalog for data and AI model governance, quality, and collaboration.



A

Account Team
The amount of time in seconds an incoming chat request remained unanswered by an agent before the chat was disconnected by the customer.
Customer Name | Customer Name Addresses | +3 more
Last modified: 24 Jul 2018

Alternate Service Identifier Inactive
Code that uniquely identifies a service (such as a given service activity) in a systems that does not identify.
Services
Last modified: 24 Jul 2018

B

Bank Account
A financial account maintained by a financial institution for a customer.
Economic and Financial
Last modified: 24 Jul 2018

Bank Account Name
The familiar name given to a specific banking account

Information Governance Domain Policies Draft

General

Description
Social media are interactive computer-mediated technologies that facilitate the creation and sharing of information, ideas, career interests and other forms of expression via virtual communities and networks.
Show more

Primary Category
None

Secondary Category
None

Attachments
None

Rules

Description
No description provided

Created by: admin
Last modification: Oct 29, 2019
Connections used: 1
All data sets: 12 View
Reviewed data sets: 0

Governed workspace
Sampling enabled

Search for terms in global search

Import new reference data from a csv file

Assess data quality

WKC Use Cases

- Metadata Management
- Regulatory compliance
- Improve Data Quality
- Enterprise Data & AI Governance
- Data monetization and self-service
- Enterprise Data Cataloging
- Data lake consumption
- Reference Data Management
- DataOps Maturity

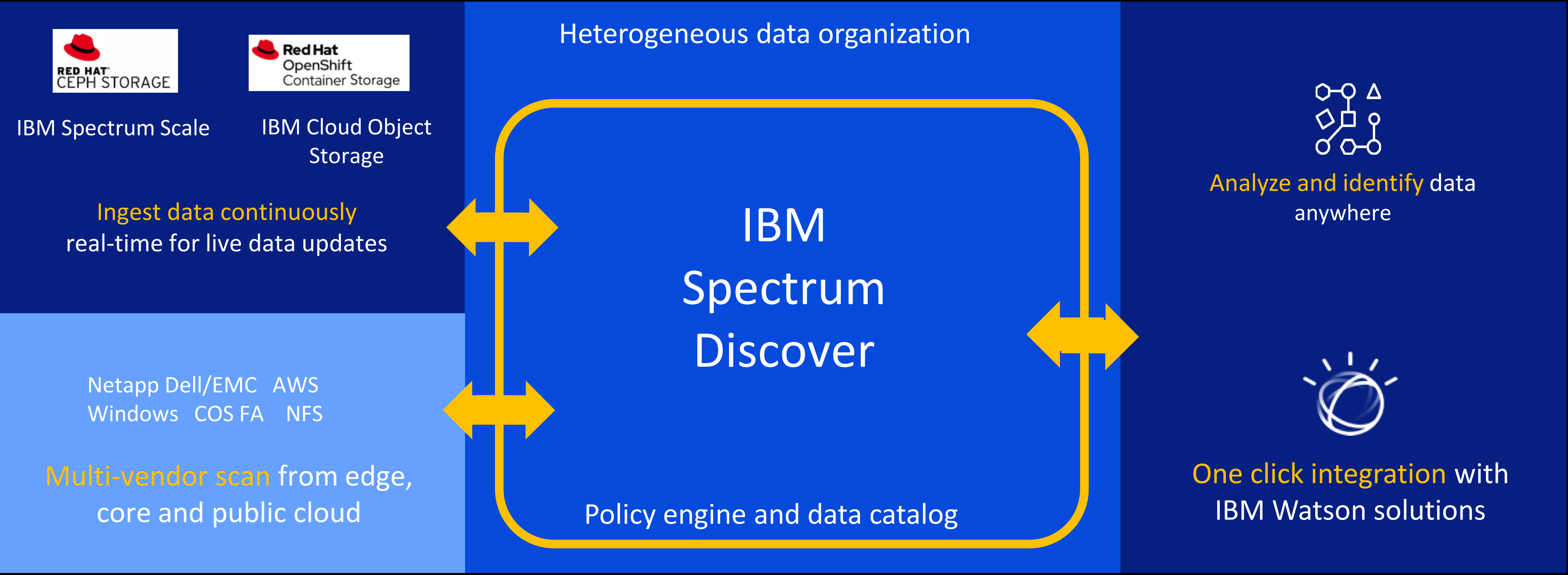
An unstructed data catalog and policy engine to organize the AI infrastructure and help solve the data and AI puzzle faster

AI Workflows	Data Curation	Data Analysis

IBM Spectrum Discover

IBM Spectrum Discover

Catalog your data and search billions of files/objects in 0.5 sec
Manage AI workflows, data security analysis and data governance



RED HAT CEPH STORAGE

Red Hat OpenShift Container Storage

IBM Spectrum Scale

IBM Cloud Object Storage

Ingest data continuously
real-time for live data updates

Netapp Dell/EMC AWS
Windows COS FA NFS

Multi-vendor scan from edge, core and public cloud

Heterogeneous data organization

IBM Spectrum Discover

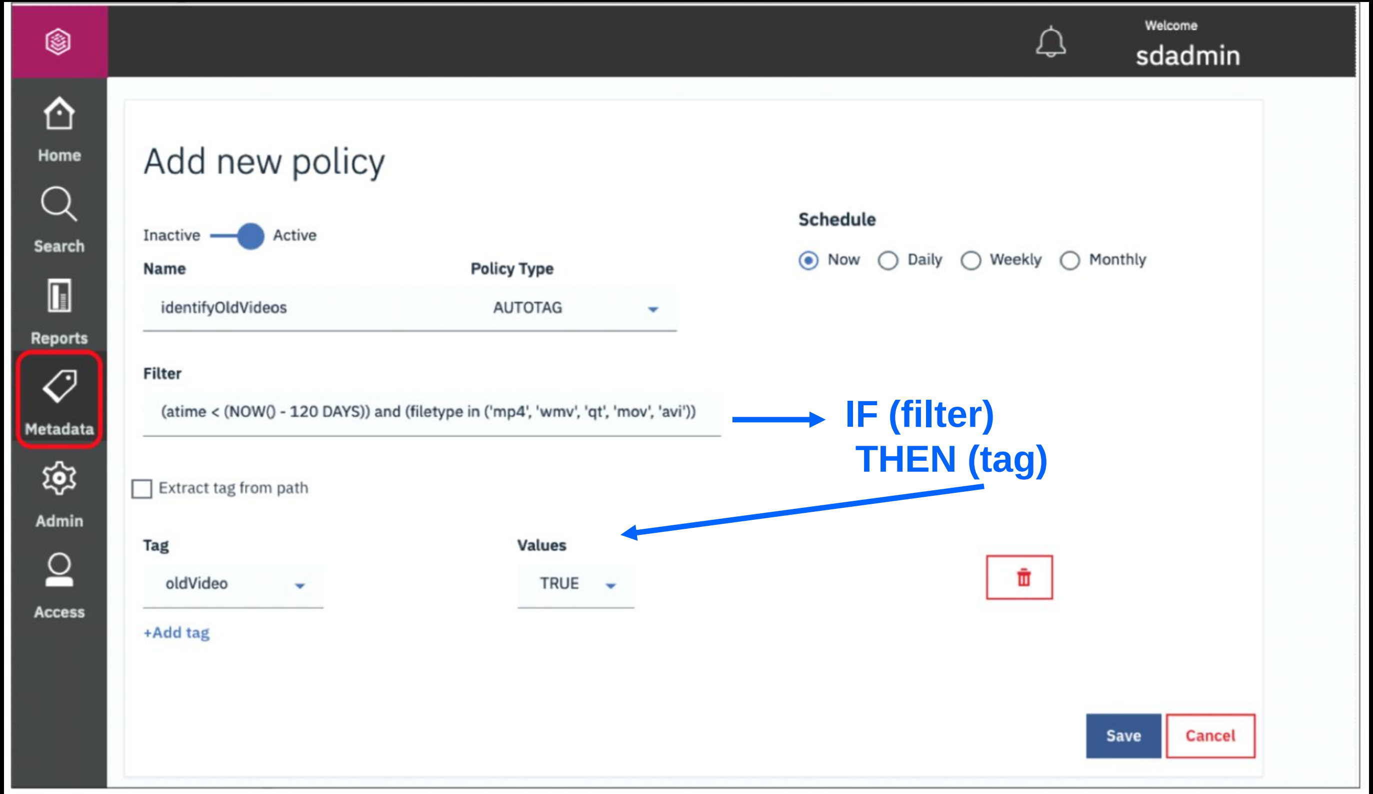
Policy engine and data catalog

Analyze and identify data anywhere

One click integration with IBM Watson solutions

Easily create a policy to “tag” items based on a filter

Data Mapping



Add new policy

Inactive ● Active

Name: identifyOldVideos

Policy Type: AUTOTAG

Schedule: ● Now ○ Daily ○ Weekly ○ Monthly

Filter: (atime < (NOW() - 120 DAYS)) and (filetype in ('mp4', 'wmv', 'qt', 'mov', 'avi'))

□ Extract tag from path

Tag: oldVideo

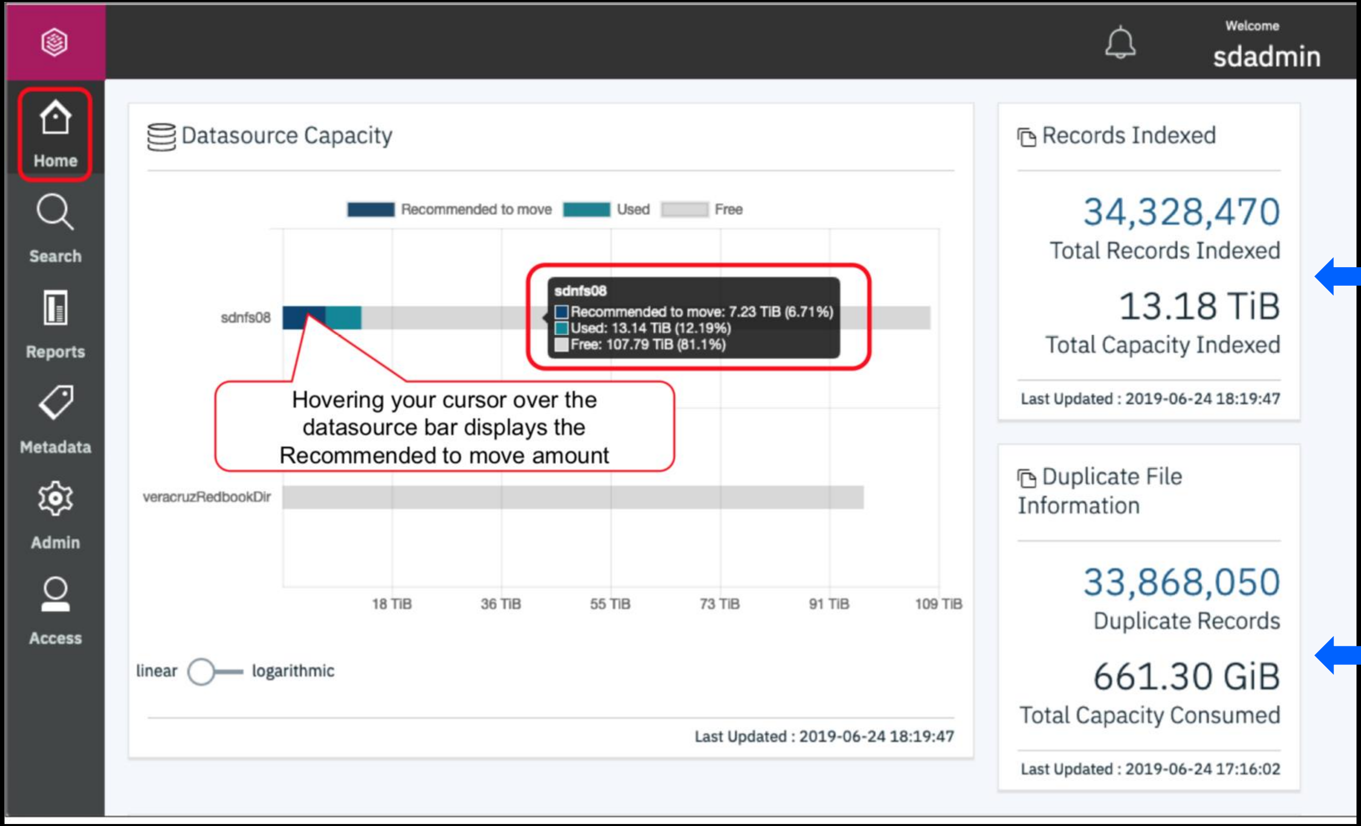
Values: TRUE

+Add tag

IF (filter)
THEN (tag)

Discover in one screen
duplicate records and data for archive

Data Visualization



Welcome sdadmin

Datasource Capacity

sdnfs08
- Recommended to move: 7.23 TiB (6.71%)
- Used: 13.14 TiB (12.19%)
- Free: 107.79 TiB (81.1%)

Hovering your cursor over the datasource bar displays the Recommended to move amount

Last Updated : 2019-06-24 18:19:47

Records Indexed

34,328,470
Total Records Indexed

13.18 TiB
Total Capacity Indexed

Last Updated : 2019-06-24 18:19:47

Duplicate File Information

33,868,050
Duplicate Records

661.30 GiB
Total Capacity Consumed

Last Updated : 2019-06-24 17:16:02

Summary of capacity

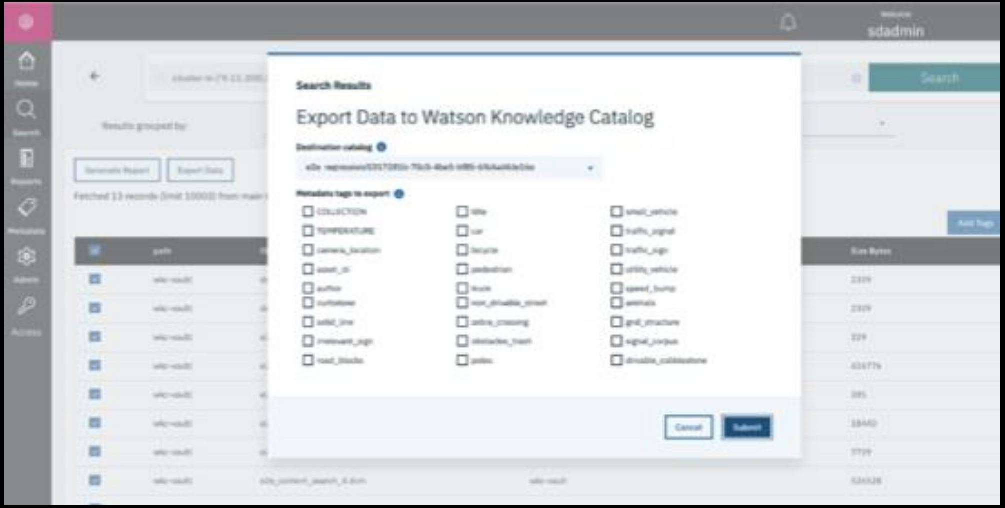
Summary of duplicate records

A simple one click export to IBM Watson Knowledge Catalog

- Automatically register assets with Watson Knowledge Catalog

- Leverage assets in IBM Cloud Pak for Data

- Import custom tags and create new and expanded insights from data



Search Results

Export Data to Watson Knowledge Catalog

Destination catalog

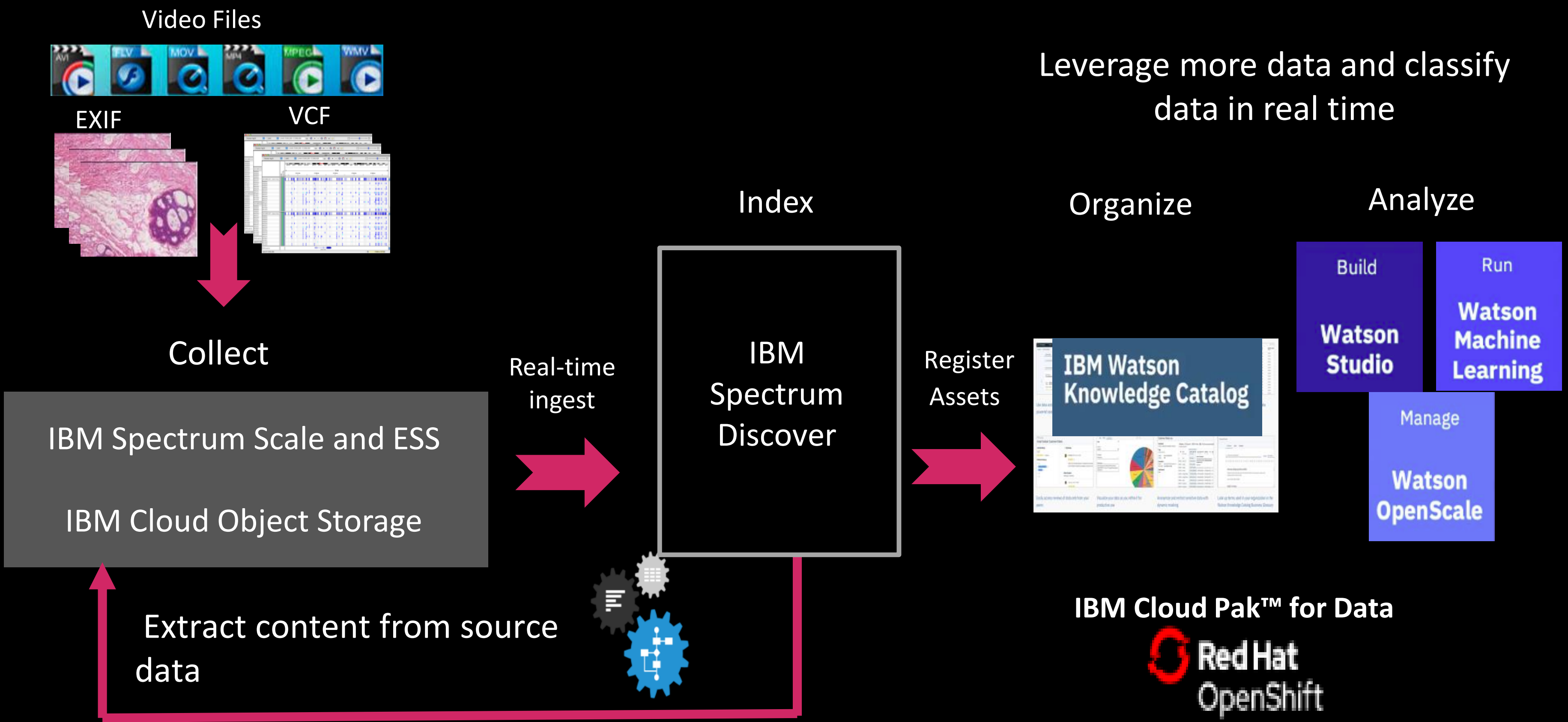
[illegible]

Metadata tags to export

□ COLLECTION
□ TEMPERATURE
□ camera_location
□ asset_id
□ author
□ curbstone
□ solid_line
□ irrelevant_sign
□ road_blocks
□ title
□ car
□ bicycle
□ pedestrian
□ truck
□ non_drivable_street
□ zebra_crossing
□ obstacles_trash
□ poles
□ small_vehicle
□ traffic_signal
□ traffic_sign
□ utility_vehicle
□ speed_bump
□ animals
□ grid_structure
□ signal_corpus
□ drivable_cobblestone

Cancel Submit

Smarter data with integrated IBM solutions



Video Files

EXIF

VCF

Collect

IBM Spectrum Scale and ESS

IBM Cloud Object Storage

Real-time ingest

Index

IBM Spectrum Discover

Extract content from source data

Register Assets

Leverage more data and classify data in real time

Organize

IBM Watson Knowledge Catalog

Analyze

Build

Watson Studio

Run

Watson Machine Learning

Manage

Watson OpenScale

IBM Cloud Pak™ for Data

Red Hat OpenShift

Learn more

Watson Knowledge Catalog at
ibm.com/watson-knowledge/catalog

IBM Spectrum Discover at
https://www.ibm.com/products/spectrum-discover

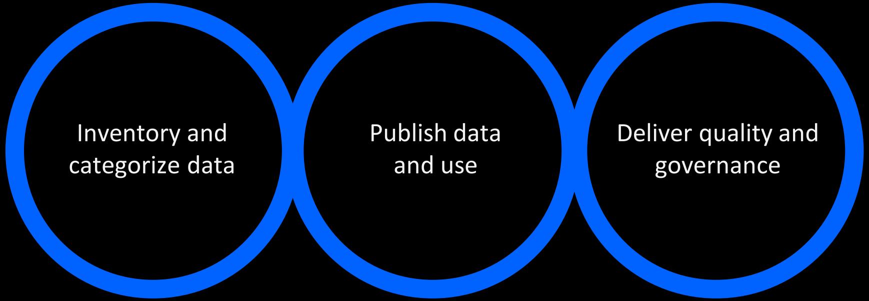
Introducing DataOps

> "DataOps is a collaborative data management practice focused on improving the communication, integration and automation of data flows between data managers and data consumers across an organization."

— Gartner

DataOps Methodology Begins with Automating Metadata Management Best Practices

Business objectives



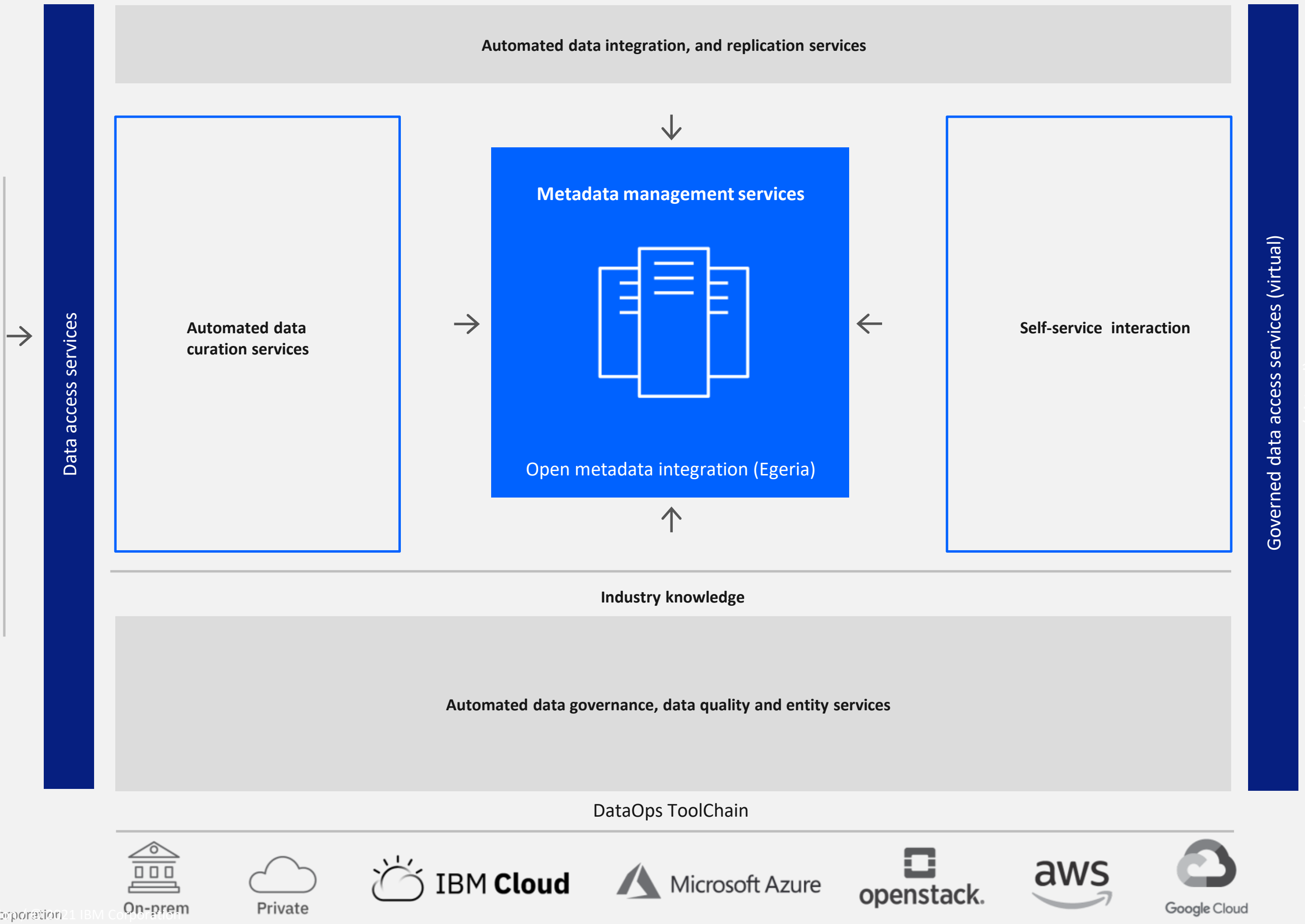
DataOps Methodology

- Align data pipelines with business objective and success criteria.
- Automatically measures accuracy and speed of data capture, quality and use.
- Automates data and metadata ingestion and classification.
- Automatically assesses data quality issues and alerts when anomalies are detected.
- Automatically initiates remediation via workflow.
- Automatically ensures authorized use of published data assets by enforcing data privacy and governance policies.

DataOps capabilities

Data Sources

- Systems of record
- IoT
- Systems of insights
- Cloud
- Hadoop
- Social media
- Unstructured
- Other external
- Logs



Data access services
Automated data integration, and replication services
Automated data curation services
Metadata management services
Open metadata integration (Egeria)
Self-service interaction
Governed data access services (virtual)
Industry knowledge
Automated data governance, data quality and entity services
DataOps ToolChain
On-prem
Private
IBM Cloud
Microsoft Azure
openstack.
aws
Google Cloud

Users

- Chief data officer
- Governance officers
- Data quality analyst
- Data steward
- Data scientist
- Business users
- Data engineer
- Application developer
- Application tester

Thank you

Legal notices

Product data has been reviewed for accuracy as of the date of initial publication. Product data is subject to change without notice. This document could include technical inaccuracies or typographical errors. IBM may make improvements and/or changes in the product(s) and/or program(s) described herein at any time without notice. Any statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only. References in this document to IBM products, programs, or services does not imply that IBM intends to make such products, programs or services available in all countries in which IBM operates or does business. Any reference to an IBM Program Product in this document is not intended to state or imply that only that program product may be used. Any functionally equivalent program, that does not infringe IBM's intellectually property rights, may be used instead.

THE INFORMATION PROVIDED IN THIS DOCUMENT IS DISTRIBUTED "AS IS" WITHOUT ANY WARRANTY, EITHER OR IMPLIED. IBM LY DISCLAIMS ANY WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE OR NONINFRINGEMENT. IBM shall have no responsibility to update this information. IBM products are warranted, if at all, according to the terms and conditions of the agreements (e.g., IBM Customer Agreement, Statement of Limited Warranty, International Program License Agreement, etc.) under which they are provided. Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products in connection with this publication and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. IBM makes no representations or warranties, ed or implied, regarding non-IBM products and services.

The provision of the information contained herein is not intended to, and does not, grant any right or license under any IBM patents or copyrights. Inquiries regarding patent or copyright licenses should be made, in writing, to:

IBM Director of Licensing
IBM Corporation
North Castle Drive
Armonk, NY 1 0504- 785
U.S.A.

Information and trademarks

IBM, the IBM logo, ibm.com, IBM System Storage, IBM Spectrum Storage, IBM Spectrum Control, IBM Spectrum Protect, IBM Spectrum Archive, IBM Spectrum Virtualize, IBM Spectrum Scale, IBM Spectrum Accelerate, Softlayer, and XIV are trademarks of International Business Machines Corp., registered in many jurisdictions worldwide. A current list of IBM trademarks is available on the Web at "Copyright and trademark information" at http://www.ibm.com/legal/copytrade.shtml

The following are trademarks or registered trademarks of other companies.
Adobe, the Adobe logo, PostScript, and the PostScript logo are either registered trademarks or trademarks of Adobe Systems Incorporated in the United States, and/or other countries.
IT Infrastructure Library is a Registered Trade Mark of AXELOS Limited.
Linear Tape-Open, LTO, the LTO Logo, Ultrium, and the Ultrium logo are trademarks of HP, IBM Corp. and Quantum in the U.S. and other countries.
Intel, Intel logo, Intel Inside, Intel Inside logo, Intel Centrino, Intel Centrino logo, Celeron, Intel Xeon, Intel SpeedStep, Itanium, and Pentium are trademarks or registered trademarks of Intel Corporation or its subsidiaries in the United States and other countries.
Linux is a registered trademark of Linus Torvalds in the United States, other countries, or both.
Microsoft, Windows, Windows NT, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.
Java and all Java-based trademarks and logos are trademarks or registered trademarks of Oracle and/or its affiliates.
Cell Broadband Engine is a trademark of Sony Computer Entertainment, Inc. in the United States, other countries, or both and is used under license therefrom.
ITIL is a Registered Trade Mark of AXELOS Limited.
UNIX is a registered trademark of The Open Group in the United States and other countries.
* All other products may be trademarks or registered trademarks of their respective companies.

Notes:
Performance is in Internal Throughput Rate (ITR) ratio based on measurements and projections using standard IBM benchmarks in a controlled environment. The actual throughput that any user will experience will vary depending upon 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 throughput improvements equivalent to the performance ratios stated here.

All customer examples cited or described in this presentation are presented as illustrations of the manner in which some customers have used IBM products and the results they may have achieved. Actual environmental costs and performance characteristics will vary depending on individual customer configurations and conditions.
This publication was produced in the United States. IBM may not offer the products, services or features discussed in this document in other countries, and the information may be subject to change without notice. Consult your local IBM business contact for information on the product or services available in your area.
All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.
Information about non-IBM products is obtained from the manufacturers of those products or their published announcements. IBM has not tested those products and cannot confirm the performance, compatibility, or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

Prices subject to change without notice. Contact your IBM representative or Business Partner for the most current pricing in your geography.

This presentation and the claims outlined in it were reviewed for compliance with US law. Adaptations of these claims for use in other geographies must be reviewed
by the local country counsel for compliance with local laws.

IBM