

Cataloging Unstructured Data in IBM Watson Knowledge Catalog with IBM Spectrum Discover

Joseph Dain

Joshua Blumert

Abeer Selim

Larry Coyne

Anil Patil

Christopher Vollmar

Flavio de Rezende, PhD

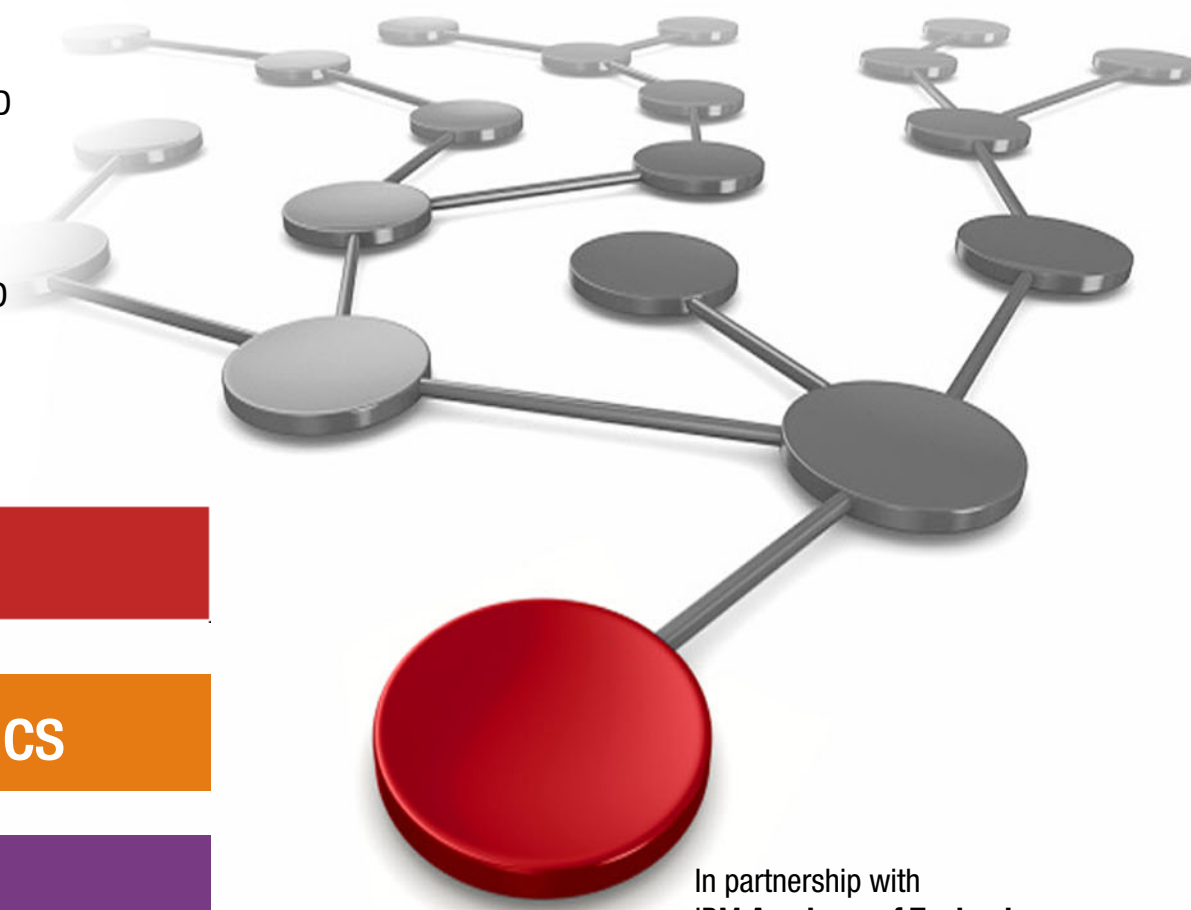
Frank Greco

Frank N. Lee, PhD

Isom Crawford Jr., PhD

Ivaylo B. Bozhinov

Joanna Wong, PhD



 **Cloud**

 **Analytics**

Storage

In partnership with
IBM Academy of Technology



IBM Redbooks

**Cataloging Unstructured Data in IBM Watson
Knowledge Catalog with IBM Spectrum Discover**

August 2020

Note: Before using this information and the product it supports, read the information in “Notices” on page v.

First Edition (August 2020)

This edition applies to Version 2, Release 0, Modification 3 of IBM Spectrum Discover (product number 5737-I32).

© Copyright International Business Machines Corporation 2020. All rights reserved.

Note to U.S. Government Users Restricted Rights -- Use, duplication or disclosure restricted by GSA ADP Schedule Contract with IBM Corp.

Contents

Notices	v
Trademarks	vi
Preface	vii
Authors	viii
Now you can become a published author, too!	x
Comments welcome	x
Stay connected to IBM Redbooks	xi
Chapter 1. IBM Spectrum Discover overview	1
1.1 Introduction	2
1.2 High-level overview	3
1.3 Major ways to use IBM Spectrum Discover	4
1.3.1 Large-scale analytics / artificial intelligence / machine learning (ML)	4
1.3.2 Data / storage optimization use case	5
1.3.3 Data governance	5
1.3.4 Data management	6
1.4 Architecture	7
1.4.1 Role-based access control	8
1.4.2 Data source connections	9
1.4.3 GUI	9
1.4.4 Reports	10
1.5 A deeper look at metadata	11
1.5.1 Cataloging metadata	11
1.5.2 Enriching metadata	11
1.5.3 Policies and user-defined metadata	12
1.5.4 IBM Spectrum Discover Application Catalog and Software Development Kit. . . .	20
1.5.5 Data movement with IBM Spectrum Discover	22
1.6 Deployment patterns	23
Chapter 2. IBM Watson Knowledge Catalog and IBM Cloud Pak for Data overview .	25
2.1 Overview of Watson Knowledge Catalog	26
2.2 Overview of IBM CP4D	27
2.2.1 IBM CP4D and WKC	27
2.3 IBM CP4D	28
Chapter 3. IBM Spectrum Discover integration with IBM Watson Knowledge Catalog architecture and benefits	31
3.1 Solution architecture	32
3.1.1 Asset registration process	32
3.2 Connecting IBM Spectrum Discover to Watson Knowledge Catalog	33
3.3 Exporting assets from IBM Spectrum Discover to Watson Knowledge Catalog	33
3.3.1 IBM Spectrum Discover tag to WKC tag mapping	34
3.4 Using assets in Watson Knowledge Catalog	35
Chapter 4. Curating unstructured data for IBM Watson Knowledge Catalog with IBM Spectrum Discover	37
4.1 Data curation workflow	38
4.1.1 Creating tags in IBM Spectrum Discover	38

4.1.2	Creating regular expressions	39
4.1.3	Creating a content inspection policy	40
4.1.4	Searching by title and author	41
4.2	Using assets in IBM CP4D and Watson Knowledge Catalog	43
4.2.1	Browsing and managing assets in a catalog	44
4.2.2	Creating projects from assets in Watson Knowledge Catalog	45
4.2.3	Creating data governance policies	48
Chapter 5.	Healthcare and life sciences use cases	49
5.1	Generic healthcare use case	50
5.1.1	IBM Spectrum Discover large-scale AI and data governance with Watson Knowledge Catalog	50
5.1.2	Data governance: Medical file classification example	52
5.1.3	Large-scale analytics, AI, and ML for healthcare and life sciences	53
5.2	COVID-19 use case	55
5.2.1	Classifying images with IBM Visual Insights	56
5.2.2	Registering assets and tags / labels into Watson Knowledge Catalog	58
5.2.3	Viewing images in Watson Knowledge Catalog	59
5.2.4	Uploading an IBM Spectrum Discover custom report into Watson Knowledge Catalog	64
5.3	Breast cancer use case	67
5.3.1	Using Data Refinery, Jupyter Notebook, or Cognos to analyze report data	71
Chapter 6.	Financial services use case: Personally Identifiable Information detection and data governance	75
6.1	Current challenges in financial industries	76
6.1.1	Customer expectations	76
6.1.2	Increasing pressure from competition	76
6.1.3	Investor expectations	76
6.1.4	Keeping up with compliance and regulations	76
6.1.5	Business agility with the latest technology	77
6.2	Protecting cardholder data with PCC DDS use case	77
6.2.1	Overview of PCI	77
6.2.2	Overview of PCI requirements	79
6.2.3	Implementing PCI DSS into business	80
6.3	Creating a data governance policy in WKC	80
6.3.1	Creating a policy	84
6.3.2	Creating rules for data protection	85
Chapter 7.	Conclusion	89
Related publications		91
IBM Redbooks		91
Other publications		91
Online resources		91
Help from IBM		92

Notices

This information was developed for products and services offered in the US. This material might be available from IBM in other languages. However, you may be required to own a copy of the product or product version in that language in order to access it.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not grant you any license to these patents. You can send license inquiries, in writing, to:

IBM Director of Licensing, IBM Corporation, North Castle Drive, MD-NC119, Armonk, NY 10504-1785, US

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some jurisdictions do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM websites are provided for convenience only and do not in any manner serve as an endorsement of those websites. The materials at those websites are not part of the materials for this IBM product and use of those websites is at your own risk.

IBM may use or distribute any of the information you provide in any way it believes appropriate without incurring any obligation to you.

The performance data and client examples cited are presented for illustrative purposes only. Actual performance results may vary depending on specific configurations and operating conditions.

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 and cannot confirm the accuracy of 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.

Statements regarding IBM's future direction or intent are subject to change or withdrawal without notice, and represent goals and objectives only.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to actual people or business enterprises is entirely coincidental.

COPYRIGHT LICENSE:

This information contains sample application programs in source language, which illustrate programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. The sample programs are provided "AS IS", without warranty of any kind. IBM shall not be liable for any damages arising out of your use of the sample programs.

Trademarks

IBM, the IBM logo, and [ibm.com](http://www.ibm.com) are trademarks or registered trademarks of International Business Machines Corporation, registered in many jurisdictions worldwide. Other product and service names might be trademarks of IBM or other companies. 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 terms are trademarks or registered trademarks of International Business Machines Corporation, and might also be trademarks or registered trademarks in other countries.

AIX®	IBM Cloud Pak®	InfoSphere®
Cognitive Business®	IBM Digital Nation™	Maximo®
Digital Nation™	IBM Elastic Storage®	Redbooks®
Global Business Services®	IBM Spectrum®	Redbooks (logo)  ®
IBM®	IBM Spectrum Storage™	System Storage™
IBM Cloud®	IBM Watson®	Watson™

The following terms are trademarks of other companies:

The registered trademark Linux® is used pursuant to a sublicense from the Linux Foundation, the exclusive licensee of Linus Torvalds, owner of the mark on a worldwide basis.

Microsoft, Windows, and the Windows logo are trademarks of Microsoft Corporation in the United States, other countries, or both.

Ceph, OpenShift, Red Hat, are trademarks or registered trademarks of Red Hat, Inc. or its subsidiaries in the United States and other countries.

VMware, and the VMware logo are registered trademarks or trademarks of VMware, Inc. or its subsidiaries in the United States and/or other jurisdictions.

Other company, product, or service names may be trademarks or service marks of others.

Preface

This IBM® Redpaper publication explains how IBM Spectrum® Discover integrates with the IBM Watson® Knowledge Catalog (WKC) component of IBM Cloud® Pak for Data (IBM CP4D) to make the enriched catalog content in IBM Spectrum Discover along with the associated data available in WKC and IBM CP4D. From an end-to-end IBM solution point of view, IBM CP4D and WKC provide state-of-the-art data governance, collaboration, and artificial intelligence (AI) and analytics tools, and IBM Spectrum Discover complements these features by adding support for unstructured data on large-scale file and object storage systems on premises and in the cloud.

Many organizations face challenges to manage unstructured data. Some challenges that companies face include:

- ▶ Pinpointing and activating relevant data for large-scale analytics, machine learning (ML) and deep learning (DL) workloads.
- ▶ Lacking the fine-grained visibility that is needed to map data to business priorities.
- ▶ Removing redundant, obsolete, and trivial (ROT) data and identifying data that can be moved to a lower-cost storage tier.
- ▶ Identifying and classifying sensitive data as it relates to various compliance mandates, such as the General Data Privacy Regulation (GDPR), Payment Card Industry Data Security Standards (PCI-DSS), and the Health Information Portability and Accountability Act (HIPAA).

This paper describes how IBM Spectrum Discover provides seamless integration of data in IBM Storage with IBM Watson Knowledge Catalog (WKC). Features include:

- ▶ Event-based cataloging and tagging of unstructured data across the enterprise.
- ▶ Automatically inspecting and classifying over 1000 unstructured data types, including genomics and imaging specific file formats.
- ▶ Automatically registering assets with WKC based on IBM Spectrum Discover search and filter criteria, and by using assets in IBM CP4D.
- ▶ Enforcing data governance policies in WKC in IBM CP4D based on insights from IBM Spectrum Discover, and using assets in IBM CP4D.

Several in-depth use cases are used that show examples of healthcare, life sciences, and financial services.

IBM Spectrum Discover integration with WKC enables storage administrators, data stewards, and data scientists to efficiently manage, classify, and gain insights from massive amounts of data. The integration improves storage economics, helps mitigate risk, and accelerates large-scale analytics to create competitive advantage and speed critical research.

Authors

This paper was produced by a team of specialists from around the world working with the IBM Redbooks team.

Joseph Dain is a Senior Technical Staff Member and Master Inventor in the IBM Systems Storage organization at Tucson, Arizona. He is on his 26th invention plateau, and has over 100 patents issued and pending worldwide. Joseph joined IBM in 2003 with a BS degree in electrical engineering, and is the Chief Architect for IBM Spectrum Discover.

Abeer Selim is an IBM Certified Experienced IT Architect, Certified Expert Specialist, and Certified Senior Solution Manager in IBM Global Business Services®. She is the Middle East and Africa CICs Custom AMS Practice Leader and MEA GBS IBM Cognitive Business® Decision Support Service Line Solution Leader. Abeer has 15 years of experience in the IT industry, and holds BS and MS degrees in biomedical and systems engineering from Cairo University in Egypt. Her speciality is in ML methodologies for brain to computer interfaces. Abeer co-authored several IEEE scientific papers. She also co-authored *Building Cognitive Applications with IBM Watson Services: Volume 4 Natural Language Classifier*, SG24-8391, and AI online courses and classroom materials for IBM Digital Nation™ Africa and Skills Academy programs. Abeer is an IBM Academy of Technology (AOT) member, and she participated in multiple AOT initiatives. Recently, Abeer was recognized as a Rockstar in the AOT initiative: Red Hat OpenShift Solution Design Guidance.

Anil Patil is a Senior Solution Manager, and Chief Architect - Cloud Application Services at IBM US. He is a Certified Cloud Architect and Cloud Solution Advisor - DevOps with more than 20 years of IT experience in Cognitive Solution, IBM Cloud, Microservices, IBM Watson API, and Cloud-Native Applications. His core experience is in Microservices, Amazon Web Services (AWS), Cloud Integration, application programming interface (API) Development, and Solution Architecture. He is Lead Solution Architect and Cloud Architect for various clients in North America. Anil is an IBM Redbooks® publication author and technical contributor for various IBM material and blogs. Anil joined IBM, US in 2013 and holds a BE degree in Electronics and an Executive MBA in finance and strategy from Rutgers Business School, US.

Christopher Vollmar is an IBM Certified Consulting IT Specialist (Level 3 Thought Leader) and Storage Architect who is based in Toronto, Ontario, Canada with the IBM Systems Group. Christopher is focused on helping customers build storage solutions by using the IBM Spectrum Storage™ Software-Defined Storage (SDS) family. He is also focused on helping customers develop private and hybrid storage cloud solutions by using the IBM Spectrum Storage family and Converged Infrastructure solutions. Christopher has worked for IBM for almost 20 years across many different areas of IBM, and has spent the past 10 years working with IBM System Storage™. Christopher holds an honours degree in political science from York University.

Flavio de Rezende, PhD is a Client Technical Leader at IBM US Public and Federal Market. He has extensive experience developing client solutions on various technologies, including databases (relational and NoSQL), enterprise content management, business intelligence, and data science. He holds a PhD degree in environmental engineering from Penn State University, with research in the area of data science that is applied to signal processing.

Frank Greco is an IBM Executive Software Architect. Frank graduated from the University of Chicago with a degree in mathematics. After an internship as an actuary with the CNA Insurance Co., he worked for IBM, where he holds a position as an Executive Software Architect focusing on AI, ML, and data science. While at IBM, he completed an MS degree in computer science at the University of Minnesota. Frank's current assignments bring him into contact with higher education institutions, the life-science/healthcare industry, and state and local governments.

Frank N. Lee, PhD is the Healthcare and Life Science industry leader for IBM Systems Group with over 20 years' of experience in scientific research and information technology. Frank's subject matter expertise (SME) started when he participated in the Human Genome Project as a research associate and bioinformatician. After joining IBM, Frank bridged into designing and deploying high-performance computing (HPC) systems in support of clients and IBM Business Partners worldwide. As an advocate for the transformation of the healthcare and life sciences industry towards precision medicine, Frank created an industry-first reference architecture; produced keynotes for dozens of conferences to promote readiness for AI; and published in IBM System Journals, IBM Redbooks publications, research papers, HPCwire editorials, and HIMSS reports.

Currently, Frank focuses on the development and deployment of high-performance data and AI (HPDA) that infuses large-scale capabilities that are deployable in hybrid/multi-cloud. On the data front, Frank leads the charge on metadata and its application for extreme-scale data management. He contributed to multiple patents on metadata and provenance management, co-led the creation of IBM Spectrum Discover software platform, and contributed to multiple AI use cases that are based on these innovations. On the cloud front, Frank leads the effort to integrate IBM software-defined infrastructure (SDI) with container orchestration platforms such as Red Hat OpenShift.

Isom Crawford Jr., PhD is a SME for SDI at IBM Washington Systems Center. He has over 20 years of experience in computer software product architecture and development. He holds a PhD degree in mathematical sciences from the University of Texas at Dallas and an MS degree in applied mathematics from Oklahoma State University. He developed and delivered multiple technical training courses, holds nine patents, and authored multiple publications, including *Software Optimization for High Performance Computing: Creating Faster Applications 1st Edition* by Wadleigh and Crawford.

Ivaylo B. Bozhinov has worked at IBM Bulgaria for 5 years as a technical support professional. His main areas of expertise are IBM Power Systems products, IBM AIX®, IBM i, and Red Hat Enterprise Linux. He has a BS degree in information technology from the State University of Librarian and Information Technology of Sofia, Bulgaria. He holds several IBM certifications, which include Hadoop Administration, Hadoop Foundation, Data Science Foundation, IBM Private Cloud, and IBM Blockchain Foundation. His areas of interest include AI, DL, ML, blockchain, and cloud.

Joanna Wong, PhD is an Executive IT Specialist with IBM Systems Client Centers. She has extensive experience in HPC application optimization and solution architecture implementation, recently focusing on software-defined solutions in life sciences. She has an AB degree in physics from Princeton University, MS and PhD degrees in physics from Cornell University, and an MBA degree from Walter Haas School of Business (University of California, Berkeley).

Joshua Blumert is an Executive IT Specialist for the IBM Public Sector Storage Engineering team. He has been with IBM for 19 years. He started as a server specialist for IBM System x focusing on Linux, VMware, and Windows systems. Joshua ran the IBM Solution Center for Financial Services in New York City before joining the storage team, where he continues as a server and application expert for distributed systems. Before joining IBM, Josh was a computational engineer for Silicon Graphics covering HPC. He holds a BS degree in physics with a focus in computer science from Rensselaer Polytechnic Institute in New York.

Larry Coyne is a Project Leader at the International Technical Support Organization, Tucson Arizona Center. He has over 35 years of IBM experience, with 23 in IBM Storage software management. He holds degrees in software engineering from the University of Texas at El Paso and project management from George Washington University. His areas of expertise include client relationship management, quality assurance, development management, and support management for IBM Storage Management Software.

Thanks to the following people for their contributions to this project:

David Wohlford
IBM CHQ, Marketing

Pallavi Galgali, Vasfi Gucer, Barry Hueston
IBM Systems

Now you can become a published author, too!

Here's an opportunity to spotlight your skills, grow your career, and become a published author—all at the same time! Join an IBM Redbooks residency project and help write a book in your area of expertise, while honing your experience using leading-edge technologies. Your efforts will help to increase product acceptance and customer satisfaction, as you expand your network of technical contacts and relationships. Residencies run from two to six weeks in length, and you can participate either in person or as a remote resident working from your home base.

Find out more about the residency program, browse the residency index, and apply online at:

ibm.com/redbooks/residencies.html

Comments welcome

Your comments are important to us!

We want our papers to be as helpful as possible. Send us your comments about this paper or other IBM Redbooks publications in one of the following ways:

- ▶ Use the online **Contact us** review Redbooks form found at:

ibm.com/redbooks

- ▶ Send your comments in an email to:

redbooks@us.ibm.com

- ▶ Mail your comments to:
IBM Corporation, IBM Redbooks
Dept. HYTD Mail Station P099
2455 South Road
Poughkeepsie, NY 12601-5400

Stay connected to IBM Redbooks

- ▶ Find us on Facebook:
<http://www.facebook.com/IBMRedbooks>
- ▶ Follow us on Twitter:
<http://twitter.com/ibmredbooks>
- ▶ Look for us on LinkedIn:
<http://www.linkedin.com/groups?home=&gid=2130806>
- ▶ Explore new Redbooks publications, residencies, and workshops with the IBM Redbooks weekly newsletter:
<https://www.redbooks.ibm.com/Redbooks.nsf/subscribe?OpenForm>
- ▶ Stay current on recent Redbooks publications with RSS Feeds:
<http://www.redbooks.ibm.com/rss.html>



IBM Spectrum Discover overview

This chapter provides a comprehensive overview of the IBM Spectrum Discover metadata management software platform. This overview helps storage administrators, data stewards, and data scientists understand the capabilities that are available to them with the addition of IBM Spectrum Discover.

This chapter includes the following topics:

- ▶ Introduction
- ▶ High-level overview
- ▶ Major ways to use IBM Spectrum Discover
- ▶ Architecture
- ▶ A deeper look at metadata
- ▶ Deployment patterns

1.1 Introduction

More than 80% of all data that is collected by organizations is not in a standard relational database. Instead, it is trapped in unstructured documents, social media posts, machine logs, images, and other data. Many organizations face significant challenges to manage this deluge of unstructured data, such as:

- ▶ Pinpointing and activating relevant data for large-scale analytics.
- ▶ Lacking the fine-grained visibility that is needed to map data to business priorities.
- ▶ Removing redundant, obsolete, and trivial (ROT) data.
- ▶ Identifying and classifying sensitive data.

IBM Spectrum Discover is a modern metadata management software that provides data insight for petabyte-scale file and object storage, storage on premises, and in the cloud. This software enables organizations to make better business decisions, and gain and maintain a competitive advantage.

The benefits of IBM Spectrum Discover are highlighted in Figure 1-1.

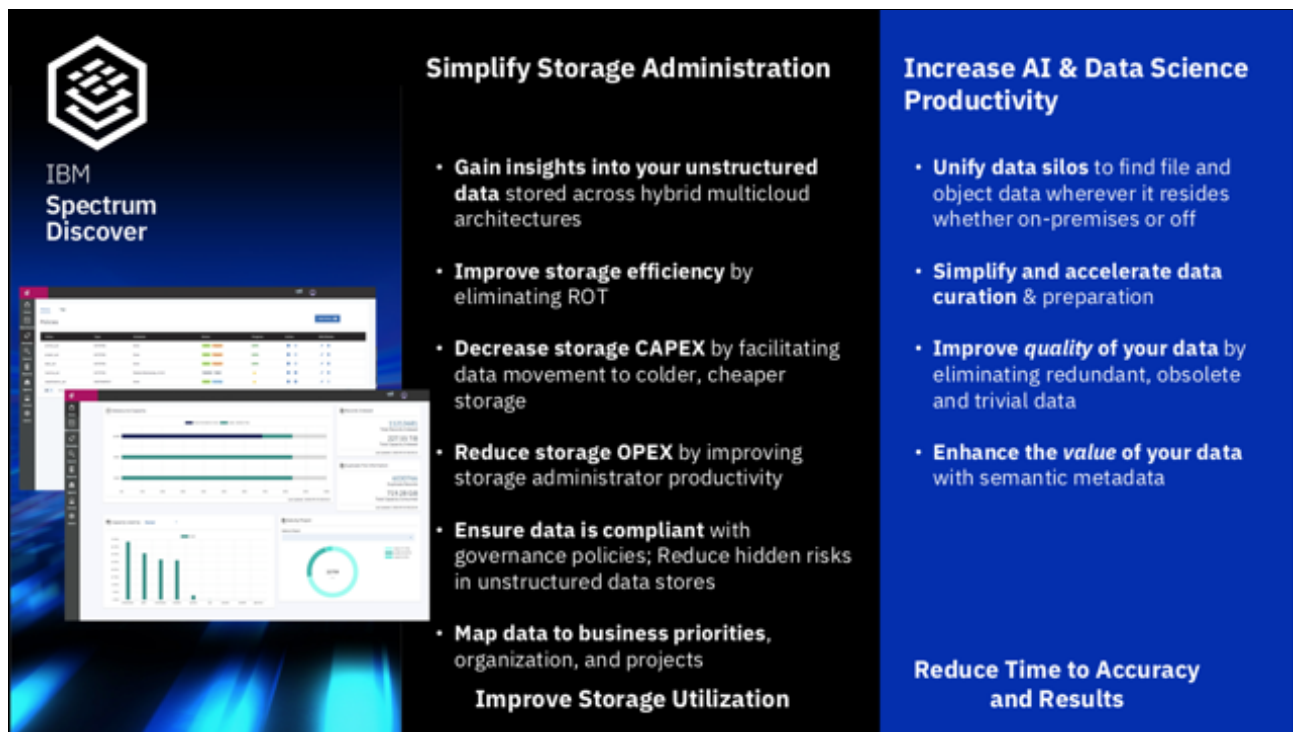


Figure 1-1 Benefits of IBM Spectrum Discover

IBM Spectrum Discover provides a rich metadata layer that enables storage administrators, data stewards, and data scientists to efficiently manage, classify, and gain insights from massive amounts of unstructured data. It also improves storage economics, helps mitigate risk, and accelerates large-scale analytics to create competitive advantage and speed-critical research.

The key capabilities of IBM Spectrum Discover are shown in Figure 1-2.

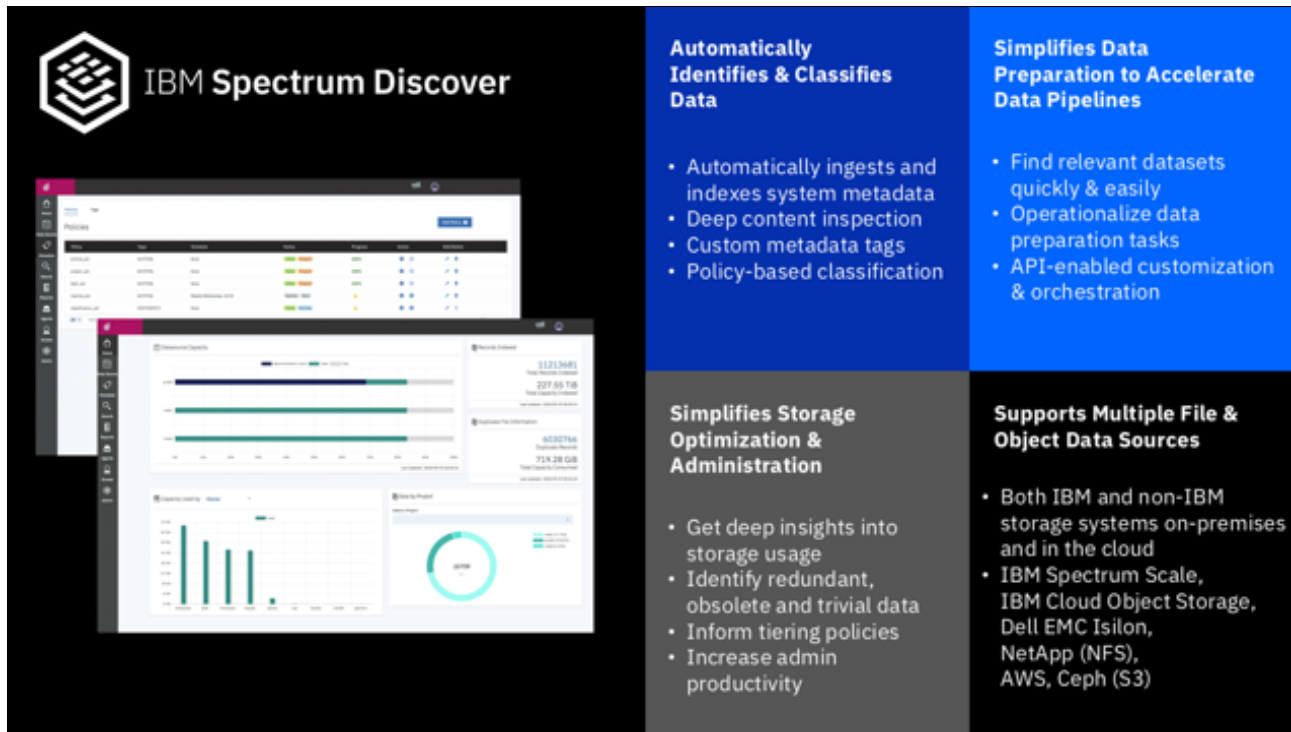


Figure 1-2 Capabilities of IBM Spectrum Discover

1.2 High-level overview

IBM Spectrum Discover is an extensible platform that provides data insight for unstructured data by scanning and cataloging metadata from storage systems. The catalog can consist of metadata from numerous storage systems, on-premises or in the cloud, which enables a holistic view of data across the entire enterprise.

Note: *Metadata* is data that describes data. Metadata captures the useful attributes of the associated source data to give the metadata context and meaning. For example, source data is a file or an object. The metadata is a set of attributes that are key-value pairs. The metadata records are associated with the file or object and are typically stored on the same system as the source data.

System metadata is created and updated by the host system and not the application software. IBM Spectrum Discover enables the addition of tags that can capture non-system metadata-specific attributes.

After the initial scan of a data source and the population of the basic metadata information within the catalog is complete, the catalog can then be enriched. The enrichment comes from more information that is derived from the internal capabilities of IBM Spectrum Discover, purpose-built applications that use the extensible platform architecture, or custom tags that act as an extension of the system metadata that can contain organizational information beyond the view and limits of the source storage system.

Note: The source storage system metadata is not relocated to IBM Spectrum Discover, but remains in the original location. The IBM Spectrum Discover catalog is populated with a pointer to the original location.

These enrichments can all be carried out from the IBM Spectrum Discover GUI and by using IBM Spectrum Discover REST application programming interfaces (APIs). Using the enriched metadata that is provided by IBM Spectrum Discover enables storage administrators and users to generate various reports that are based on any of the information that is within the catalog. Also, the enriched metadata can be used with the extensible platform architecture to perform actions on selected data.

1.3 Major ways to use IBM Spectrum Discover

This section provides an overview of the key use cases for IBM Spectrum Discover, as highlighted in Figure 1-3.

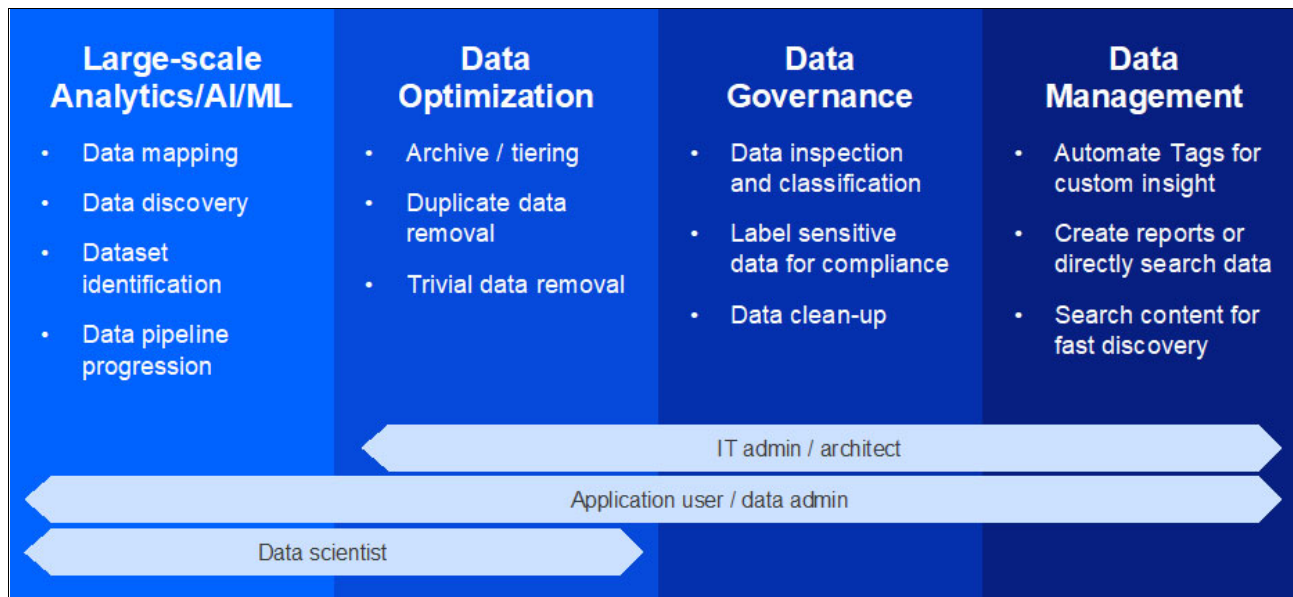


Figure 1-3 Key use cases of IBM Spectrum Discover

1.3.1 Large-scale analytics / artificial intelligence / machine learning (ML)

This use case enables data scientists and researchers to use the insights that are obtained by IBM Spectrum Discover to quickly and easily identify the relevant data for their experiment across billions of files and objects in the heterogeneous storage environment. Researchers and data scientists use either the IBM Spectrum Discover GUI or REST APIs to perform data discovery based on the information in the IBM Spectrum Discover catalog.

The researchers or data scientists can identify, classify, and organize data by applying user-defined tags that are based on the system metadata (such as file path information). Researchers and data scientists can also set up policies to perform automated content-based keyword search extract key aspects from the data and index it into IBM Spectrum Discover.

After the data set is identified, it can be automatically registered as a data set in an upstream artificial intelligence (AI) / analytics application, such as IBM Cloud Pak® for Data and IBM Watson Knowledge Catalog (WKC), which are described in Chapter 2, “IBM Watson Knowledge Catalog and IBM Cloud Pak for Data overview” on page 25 and Chapter 3, “IBM Spectrum Discover integration with IBM Watson Knowledge Catalog architecture and benefits” on page 31, and other items like IBM Watson Machine Learning Accelerator, IBM Maximo® Visual Inspection (formerly called IBM Visual Insights), or TensorFlow. The identified data set can also be automatically moved from a “warm” tier to a “hot” tier, such as Non-Volatile Memory Express (NVMe) -backed storage that is attached to GPUs through InfiniBand before starting the data set registration.

Another key aspect of this use case is creating fully automated inferencing pipelines that use the metadata event-driven architecture that is described in Chapter 3, “IBM Spectrum Discover integration with IBM Watson Knowledge Catalog architecture and benefits” on page 31 to automatically perform an inference operation on new data in the data lake and insert the resulting label information into the IBM Spectrum Discover catalog.

1.3.2 Data / storage optimization use case

This use case is understood and popular in the market because it pertains to using the system metadata that is collected across the heterogeneous storage environment to simplify storage administration and improve storage utilization by providing the following capabilities:

- ▶ Gain insights into your unstructured data that is stored across hybrid multi-cloud architectures.
- ▶ Improve storage efficiency by eliminating ROT data. Identify data by file type, size, and potentially duplicate data.
- ▶ Decrease storage CAPEX by facilitating data movement to colder, cheaper storage. Understand how your data is aging.
- ▶ Reduce storage OPEX by improving storage administrator productivity.
- ▶ Ensure that data is compliant with governance policies. Reduce hidden risks in unstructured data stores.
- ▶ Map data to business priorities, organization, and projects.

These capabilities are accomplished by the built-in analytics capabilities of the IBM Spectrum Discover platform.

1.3.3 Data governance

IBM Spectrum Discover can classify data based on user-defined tags and the content of the data. The built-in Content Classification capability of IBM Spectrum Discover supports over 1000 different file and object types that use an open source tool called Apache Tika. Apache Tika is also embedded in some IBM Watson Natural Language Processing (NLP) products to perform a similar capability. IBM Spectrum Discover also provides built-in support for Digital Imaging and Communications in Medicine (DICOM) medical images and genomics VCF data.

The Content inspection capability can identify key fields such as Social Security Number (SSN), phone numbers, account numbers, and many other fields to identify and tag content that contains Personally Identifiable Information (PII) and sensitive data by providing the following capabilities:

- ▶ Automate the identification and classification of documents that might potentially contain PII and sensitive data.
- ▶ Support for content-based data classification enables users to set up policies to automatically identify, classify, and categorize data, which can be used for specific business needs.

For the data steward and the chief information office (CIO), the ability to find and organize documents based on content greatly helps with their data administration efforts, for example, identifying data that might be subject to specific governance policies or compliance regulations.

Note: Enforcing data governance policies is not supported in IBM Spectrum Discover. Integration with the WKC component in IBM Cloud Pak for Data (IBM CP4D) supports this capability.

This use case is explored in more detail in Chapter 6, “Financial services use case: Personally Identifiable Information detection and data governance” on page 75.

1.3.4 Data management

IBM Spectrum Discover offers enhanced unstructured data management capabilities that include automatically tagging data for custom insight, creating reports or directly searching data, searching content for fast discovery, and using these insights for data movement and data cleanup. With the ability to automatically add tags to data that is based on institutional knowledge or derived information that is based on the metadata or content of the data, Data Stewards and Data Administrators can develop a much deeper understanding of the data that they need to manage.

IBM Spectrum Discover can dynamically visualize or generate reports that are based on the metadata and associated tags of the files. So, users can link data sets that might not have been previously associated, or search those data sets for a deeper understanding of what that data is. It also can run searches in the data content for attributes that are relevant to the business.

IBM Spectrum Discover provides both a search bar and a more advanced search pane to help users quickly find subsets of records that are indexed. Search results are displayed in a columnar table that contains information that is correlated to search criteria. What a user can see or not see is determined by using role-based access control (RBAC). All these features enable a strong decision-making capability around the management of the data by using the metadata about the data or data from the content files in a simple searchable format.

1.4 Architecture

IBM Spectrum Discover is an extensible platform that provides exabyte-scale data ingestion, data visualization, data activation, and business-oriented data mapping from across the enterprise. It enables data management at a broader level, which enables a more precise view of the “who, what, where, when, and why” aspects of an organization’s data rather than the myopic view of data from a single storage system. The IBM Spectrum Discover architecture is shown in Figure 1-4.

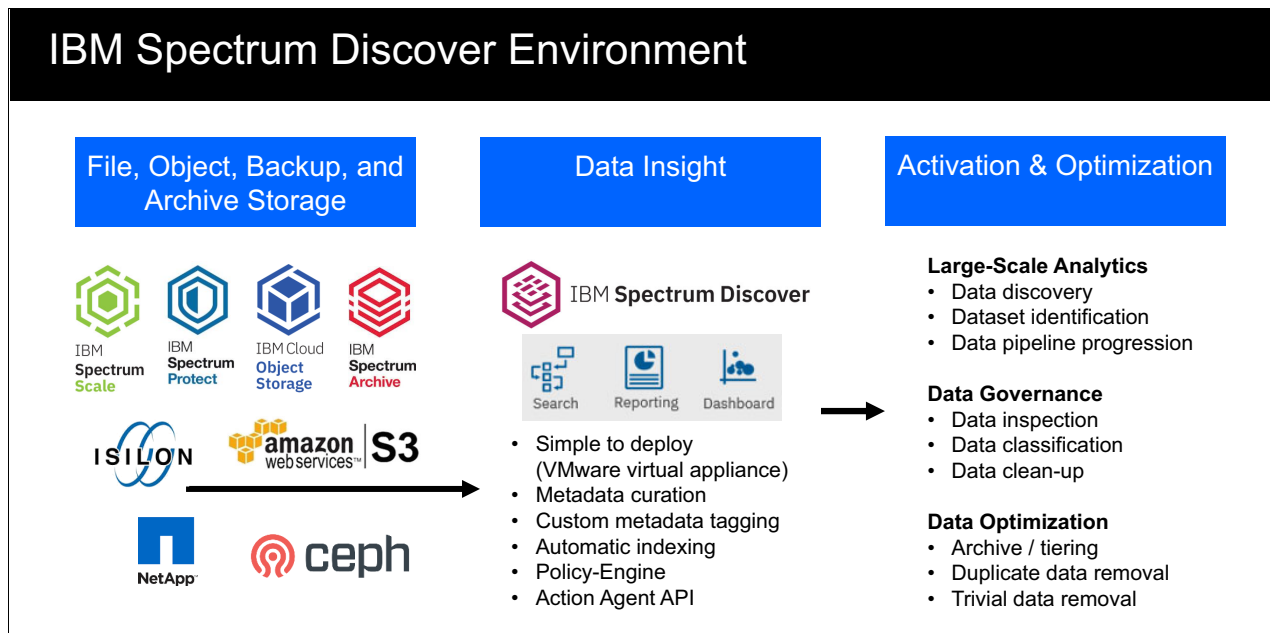


Figure 1-4 IBM Spectrum Discover architecture

IBM Spectrum Discover can scan or ingest billions of records in the course of a day. Ingesting data consists of reading metadata information from the source storage system and automatically cataloging the information into the IBM Spectrum Discover platform. This feature enables IBM Spectrum Discover to deliver results of complex queries or multi-faceted searches against the metadata information ultrafast, even when the catalog contains billions of entries. The search results are visualized by the GUI’s drill-down dashboard nearly instantaneously.

IBM Spectrum Discover easily connects to the following data sources to rapidly ingest, consolidate, and index metadata for billions of files and objects:

- ▶ IBM Cloud Object Storage (IBM COS)
- ▶ IBM Spectrum Scale and IBM Elastic Storage® Server
- ▶ IBM Spectrum Protect
- ▶ IBM Spectrum Archive
- ▶ Isilon Network File System (NFS) exports and Server Message Block (SMB) shares
- ▶ NetApp NFS exports and SMB shares
- ▶ Amazon Web Services (AWS) Simple Storage Service (S3)
- ▶ Red Hat Ceph S3

IBM Spectrum Discover provides a rich metadata layer that enables storage administrators, data stewards, and data scientists to efficiently manage, classify, and gain insights from massive amounts of unstructured data. It improves storage economics, helps mitigate risk, and accelerates large-scale analytics to create competitive advantage and speed critical research.

IBM Spectrum Discover is extensible, which provides a mechanism for communication with applications that can provide even greater insight into selected data by interrogating the contents of the full data, rather than just the metadata.

The IBM Spectrum Discover platform embeds an Apache Kafka instance, which enables a communication stream that can publish and subscribe to streams of records, similar to an enterprise messaging system. The streams of records are processed as they occur.

This feature enables IBM Spectrum Discover to enhance the contents of the catalog with the results of the inspection of the data when they become available. However, by using this same mechanism, real-time streaming data pipelines reliably get data between systems or applications that can enable even more capabilities, such as moving or deleting data. In addition to reliability, this extensible design provides an amazing amount of flexibility, so the possible use cases for IBM Spectrum Discover are nearly limitless.

To capitalize on the flexible, extensible architecture of IBM Spectrum Discover, the following APIs are provided:

- ▶ Policy management API
- ▶ Custom application API

The custom application API is used to establish the Kafka topic interfaces that are used for messaging and carrying out work to be done on selected data, hence the name *action agent*.

The policy management API is a RESTful web service that creates, lists, updates, and deletes policies. The policy management API also provides the means to start a policy immediately or schedule it to run on a schedule.

Business-oriented data mapping can be carried out by the policy management API. An example of business-oriented data mapping is adding a project name to the catalog that is based on the location (or path) to the data.

1.4.1 Role-based access control

IBM Spectrum Discover provides access to resources that are based on roles. Customers can restrict access to information based on roles. The role that is assigned to a user or group determines their privileges.

Users and groups can be associated with collections, which use policies that determine the metadata that is available to view. User and group access can be authenticated by IBM Spectrum Discover, an LDAP server, or the IBM COS System. The administrator can manage the user access functions.

Roles

Roles determine how users and groups access records or the IBM Spectrum Discover environment. If a user or group is assigned to multiple roles, the least restrictive role is applicable. For example, if you are assigned the role of Data User and you are also assigned the role of a Data Admin, you have the privileges of a Data Admin.

The following roles are available:

Admin	This role can create users, groups, and collections, and manage LDAP connections. This role can use the Application Management APIs to install, upgrade, or delete IBM Spectrum Discover applications that use the IBM Spectrum Discover API service.
Data Admin	This role can access all metadata that is collected by IBM Spectrum Discover, and is not restricted by policies or collections. This role can also define tags and policies, including policies that assign a collection value to a set of records.
Collection Admin	Manages data within a list of collections and user and group access to those collections. The Collection Admin role is a bridge between the Data Admin role and the Data User role. Users with the Collection Admin role can create, update, and delete the policies for the collections that they administer and view, update, and delete policies of data users for the collections they administer.
Data User	This role can access metadata that is collected by IBM Spectrum Discover, but metadata access can be restricted by policies in the collections that are assigned to users in this role. This role can also define tags and policies, based on the collections to which the role is assigned.
Service User	IBM service and support personnel.

1.4.2 Data source connections

A data source connection specifies the parameters for cataloging metadata from a source system to IBM Spectrum Discover.

Without the proper connection information, ingesting metadata from a connected system fails. You can use the data source connections page of the GUI to view connection information for the data sources that are connected to your environment.

1.4.3 GUI

The IBM Spectrum Discover GUI is a portal that is used for administration purposes and running data searches, report generation, policy and tag management, and user Access Management.

Note: Based on a user's role, they might not have access to all areas of the GUI.

The IBM Spectrum Discover GUI Dashboard is shown in Figure 1-5.

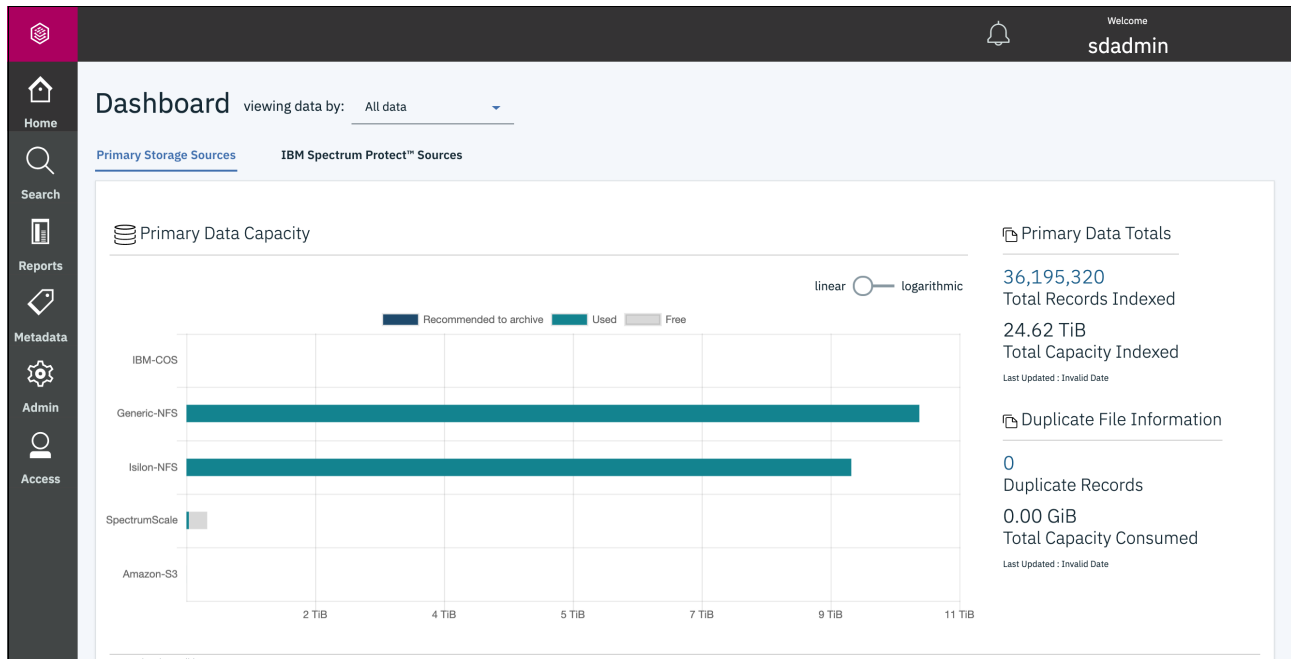


Figure 1-5 IBM Spectrum Discover GUI Dashboard

Understanding size and capacity differences

IBM Spectrum Discover collects size and capacity information.

Consider the following points:

- ▶ *Size* refers to the size of a file or object in bytes.
- ▶ *Capacity* refers to the amount of space the file or object is using on the source storage in bytes.

For objects, size and capacity values always match. However, for files, the size and capacity values can be different because of file system block overhead or sparsely populated files.

Note: Storage protection overhead (such as RAID values or erasure coding) and replication overhead are not captured in the capacity values.

1.4.4 Reports

Reports for IBM Spectrum Discover are grouped or non-grouped:

- ▶ *Grouped* reports feature information for count and sum in columns.
- ▶ *Non-grouped* reports feature information in rows.

Data Curation Reports are a way for administrators to view the state of their storage environment in different ways. They can range from high-level grouped information to individual record-level information.

For example, you can sort a report by owner, project, and department, or you can generate a list of records that meet specific criteria.

For more information about generating reports, see [IBM Knowledge Center](#).

1.5 A deeper look at metadata

Metadata is at the heart of IBM Spectrum Discover (collecting, creating, analyzing, reporting, and other activities). Metadata collection and creation involve the use of tags, policies, and agents. Management and exploration of metadata also involves tags and policies, and searching, visualization, and reporting.

IBM Spectrum Discover users can collect, define, explore, and report metadata. In this section, all these topics are described so that you can quickly develop a working knowledge of IBM Spectrum Discover and its essential core capabilities.

1.5.1 Cataloging metadata

Cataloging metadata in IBM Spectrum Discover is the process of ingesting and indexing the system metadata records from a source. Cataloging metadata transforms the metadata records into data that the user can reference and use.

The system metadata collection depends upon what type of storage system is being scanned. File system scans result in a different set of system metadata than object storage systems. File system metadata include tags, such as size, owner, path, file name, access time (atime), and modification time (mtime). In contrast, object storage system metadata includes bucket name, object name, object length, content type, system UUID, and other data.

IBM Spectrum Discover features live event notifications for IBM COS data connections. These notifications are triggered by user actions on the source data and result in updating the system metadata on the IBM Spectrum Discover system. Example actions include reading, writing, moving, and deleting data, and changing permissions or ownership. The events generate a metadata record in real time that is stored in IBM Spectrum Discover.

1.5.2 Enriching metadata

IBM Spectrum Discover can enrich the metadata from supported platforms with more information by using custom tags, policies, and action agents.

Tags

A *tag* is a custom metadata field (or key-value pair) that is used to supplement storage system metadata with organization-specific information.

An organization might segment their storage by project or by chargeback department. Those facets do not show in the system metadata, and the storage systems do not provide management and reporting capabilities based on those organizational concepts. By using custom tags, you can store more information, and manage, report, and search for data by using that organizationally important information.

Types of tags

The following types of tags are available:

- ▶ **Categorization tags:** Contain values such as project, department, and security classification. Categorization tags can be open or restricted. If it is open, listed selections can be used. If it is closed, selection is limited to true or false.
- ▶ **Characteristic tags:** Can contain any value that is needed to describe or classify the object. Can contain long descriptive values. Size limit is 4 KB.

1.5.3 Policies and user-defined metadata

One of the most powerful aspects of IBM Spectrum Discover is its ability to accommodate metadata that is created by users of the associated data or the IBM Spectrum Discover system. You can collect metadata by logically combining metadata, using header extraction tools (such as Apache Tika), or defining your own software agents to extract metadata unique to your business.

Policies are used to add information about the source data that is indexed in IBM Spectrum Discover. A policy determines the set of files to add tag values to or send to an action agent through filtering criteria. The policies give the user the ability to run actions one time or on a set schedule. Policies work in batches and can be paused, resumed, stopped, and restarted. The user can control the load on the IBM Spectrum Discover system and on the source storage system in the case of deep inspection policies.

A policy includes the following components:

Policy ID	Name of the policy.
Filter	Selects a set of documents to work.
Action	ID, parameters, and schedule.

Three basic categories of policies support user-defined metadata collection policies:

- ▶ AUTOTAG
- ▶ CONTENT SEARCH
- ▶ DEEP-INSPECT

The following sections describe creating these policies. Tags are a prerequisite for policies because the purpose of a policy is to assign values to one or more tags.

Policies feature a few common attributes, such as the name of the policy, the filter that is associated with the policy, and one or more tags to which the policy assigns values. When defining a filter, you use the same syntax as the **WHERE** clause in a standard SQL query. Ultimately, the filter identifies a subset of objects to which the policy applies.

The first step for any policy is to create it.

A policy tags a set of records based on filter criteria with a predefined set of tags. The ease of creating an AUTOTAG policy by using the IBM Spectrum Discover GUI is shown in Figure 1-6 on page 13.

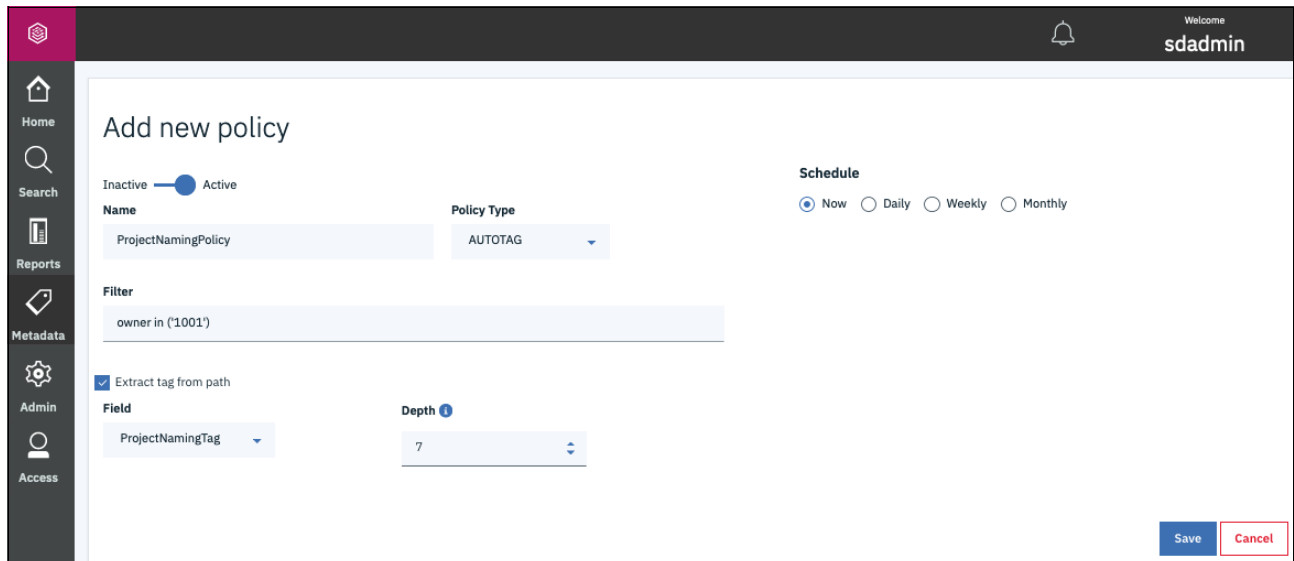


Figure 1-6 AUTOTAG policy

AUTOTAG

In an example, we define the restricted tag **oldVideo** to have only values of TRUE and FALSE. Select the **Metadata** icon in left pane of the IBM Spectrum Discover GUI and select **Policies tab** → **Add Policy**. In the resulting “Add new policy” window, complete the following steps:

1. Choose and enter the name for the policy.
2. Select **AUTOTAG** as the **Policy Type**.
3. Define the criteria for assigning the tag the wanted value, that is, define the **Filter** for the policy.
4. Select **Add tag**.
5. Select the wanted tag by using the **Tag** drop-down list.
6. Select the wanted value by using the **Values** drop-down list.

The AUTOTAG policy is basic in its operation. From a logical perspective, it can be represented as shown in the following example:

```
if (<filter>) then <tag> = <value>
```

For example, consider the defined policy that is shown in Figure 1-7. In this example, the policy `identifyOldVideos` is an AUTOTAG policy that assigns a value of TRUE to the tag `oldVideo` if the file satisfies the condition (filter):

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

For example, if the file was not accessed in 120 days and is a video file (that is, it has a file extension of `.mp4`, `.wmv`, `.qt`, `.mov`, or `.avi`), set the file's `oldVideo` tag to value of TRUE.

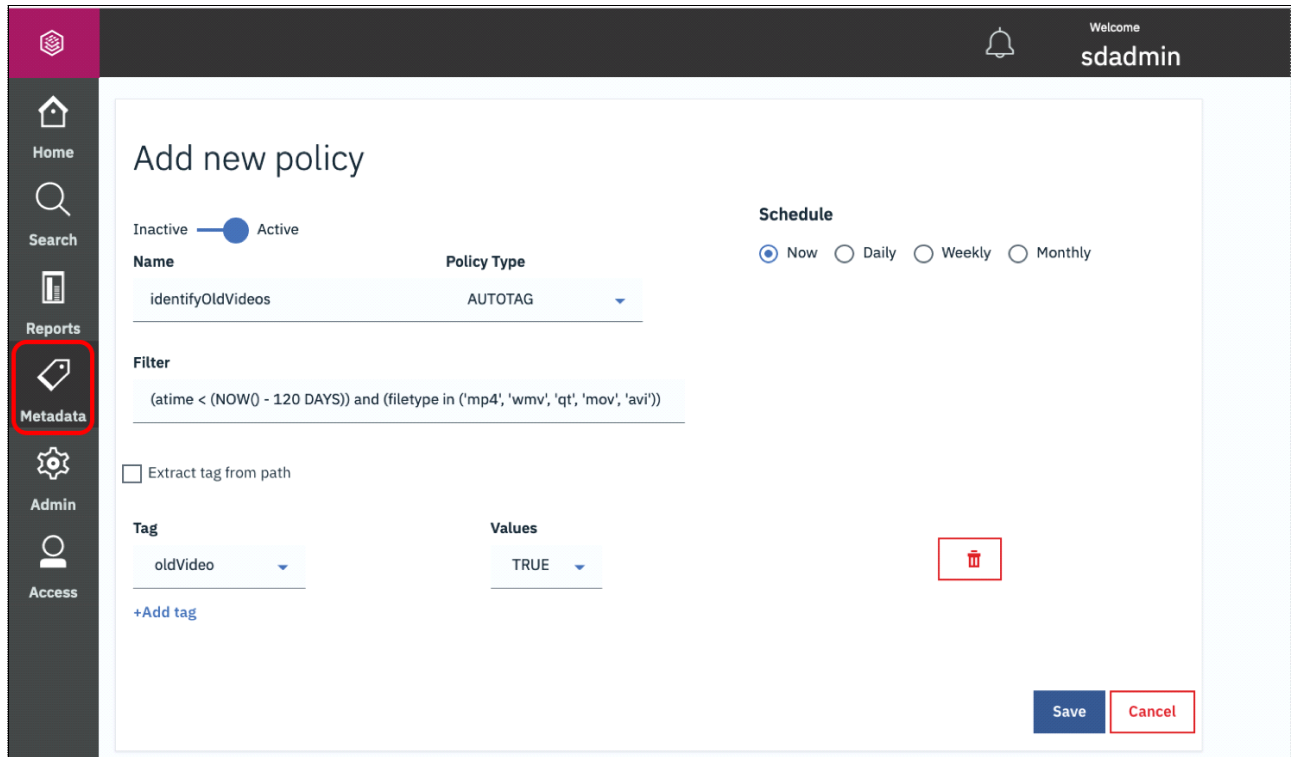


Figure 1-7 Defining an AUTOTAG policy with IBM Spectrum Discover

CONTENT SEARCH

Beginning with Version 2.0.1, IBM Spectrum Discover can enrich metadata through content inspection of source data by using the built-in CONTENTSEARCH agent. To use this function, you define regular expressions (regex) to search for and create policies that use these regex.

When the policy runs, the files or objects are retrieved from the source system by the CONTENTSEARCH agent, converted to text format if necessary, and searched by using the defined regex. The results of the search are returned to IBM Spectrum Discover and the metadata of the files that are updated according to the policy's definition.

Regular expressions,

Before defining your CONTENTSEARCH policy, identify the regex that you want to use with the policy. To do so, select the **Metadata** icon on the left pane of the IBM Spectrum Discover GUI and then select the **Regular Expressions** tab on the Metadata page. The list of available regex as displayed by the GUI is shown in Figure 1-8 on page 15. Search through the list for expressions that apply to your scenario.

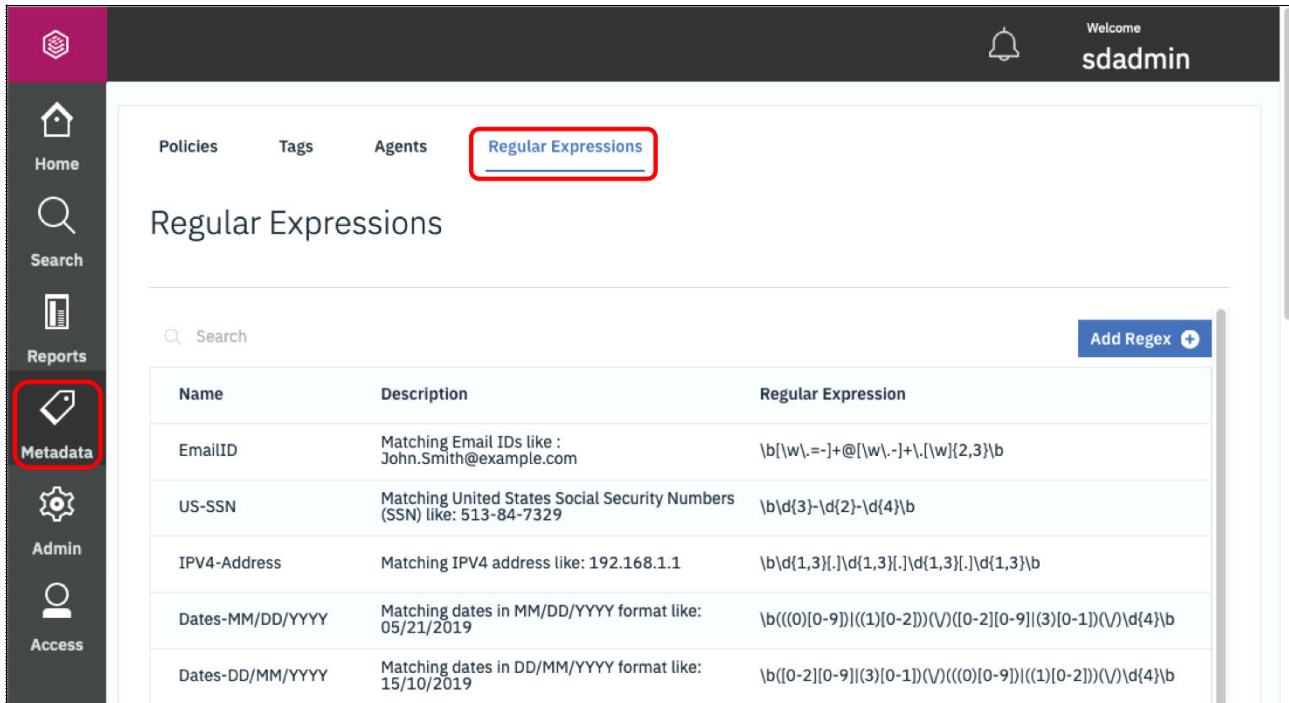


Figure 1-8 IBM Spectrum Discover provides a predefined list of regular expressions

Table 1-1 lists the IBM Spectrum Discover V2.0.3 Regular Expressions that are included with the base installation.

Table 1-1 Regular Expressions that are included with the base installation

Name	Description	Regular Expression (Pattern)
VisaCard	Matching Visa Card numbers like: 4563-7568-5698-4587	<code>\b([4]\d{3}[\s]\d{4}[\s]\d{4}[\s]\d{4}) ([4]\d{3}[\s]\d{4}[\s]\d{4}[\s]\d{4}) ([4]\d{3}[\s]\d{4}[\s]\d{4}[\s]\d{4}) ([4]\d{3}[\s]\d{4}[\s]\d{4}[\s]\d{4})\b</code>
AmexCard	Matching American Express Card numbers like: 340000000000009	<code>\b3[47][0-9]{13}\b</code>
MasterCard	Matching MasterCard numbers like: 5258704108753590	<code>\b(?:5[1-5][0-9]{2} 222[1-9] 22[3-9][0-9] 2[3-6][0-9]{2} 27[01][0-9] 2720)[0-9]{12}\b</code>
USZIPCode	Matching United States ZIP codes like: 97580	<code>\b((\d{5}-\d{4}) (\d{5})) ([A-Z]\d[A-Z]\s\d[A-Z]\d)\b</code>
URL	Matching URLs like: http://www.test.com/dir/filename.jpg?var1=foo#bar&	<code>\b((http[s]? ftp):V)?V?([^\s]+)((\w+)*V)([\w\-\.] ^#\s+)(.*)?#[\w\-\.]?#\b</code>
EmailID	Matching Email IDs like: John.Smith@example.com	<code>\b[\w\.-]+@[w\.-]+\.[w]{2,3}\b</code>
US-SSN	Matching United States like: 513-84-7329	<code>\b\d{3}-\d{2}-\d{4}\b</code>
IPV4-Address	Matching IPV4 address like: 192.168.1.1	<code>\b\d{1,3}[\.]\d{1,3}[\.]\d{1,3}[\.]\d{1,3}\b</code>
Dates-MM/DD/YYYY	Matching dates in MM/DD/YYYY format like: 05/21/2019	<code>\b(((0)[0-9]) ((1)[0-2]) (V) ([0-2][0-9]) (3)[0-1]) (V)\d{4})\b</code>
Dates-DD/MM/YYYY	Matching dates in DD/MM/YYYY format like: 15/10/2019	<code>\b([0-2][0-9]) (3)[0-1]) (V) (((0)[0-9]) ((1)[0-2]) (V)\d{4})\b</code>

Name	Description	Regular Expression (Pattern)
Currency	Matching currency like: 123, 25.50	<code>\b(d+(\.d{2})?)\b</code>
CVV-Number	Matching Credit Card Verification value number like: 670, 0927	<code>\b([0-9]{3,4})\b</code>
CreditCardExpirationDate	Matching Credit Card Expiration Date like: 11/12	<code>\bd{2}\d{2}\b</code>
CanadianSIN	Matching Canadian Social Insurance Number like: 123-456-789	<code>\b(d{3}[s]d{3}[s]+d{3})\b\b(d{3}[-]d{3}[-]+d{3})\b</code>
Geo-Coordinate	Matching Geo-Coordinates like: 51.498134, -0.201755	<code>\b([+]?)([d]{1,2})(((\d+)(,)))(s*)(([-+]?)([d]{1,3})((\d+)?))\b</code>

For more information, see [IBM Knowledge Center](#).

This basic list of regex might meet your needs. If it does not meet your needs, you can create a regex by selecting the **Add Regex** option, which opens a window in which you define a regex to be used by IBM Spectrum Discover. Enter a suitable name, description, and the regex pattern. Many useful regex and tutorials for creating them are available.

Figure 1-9 shows an example of creating a regex.

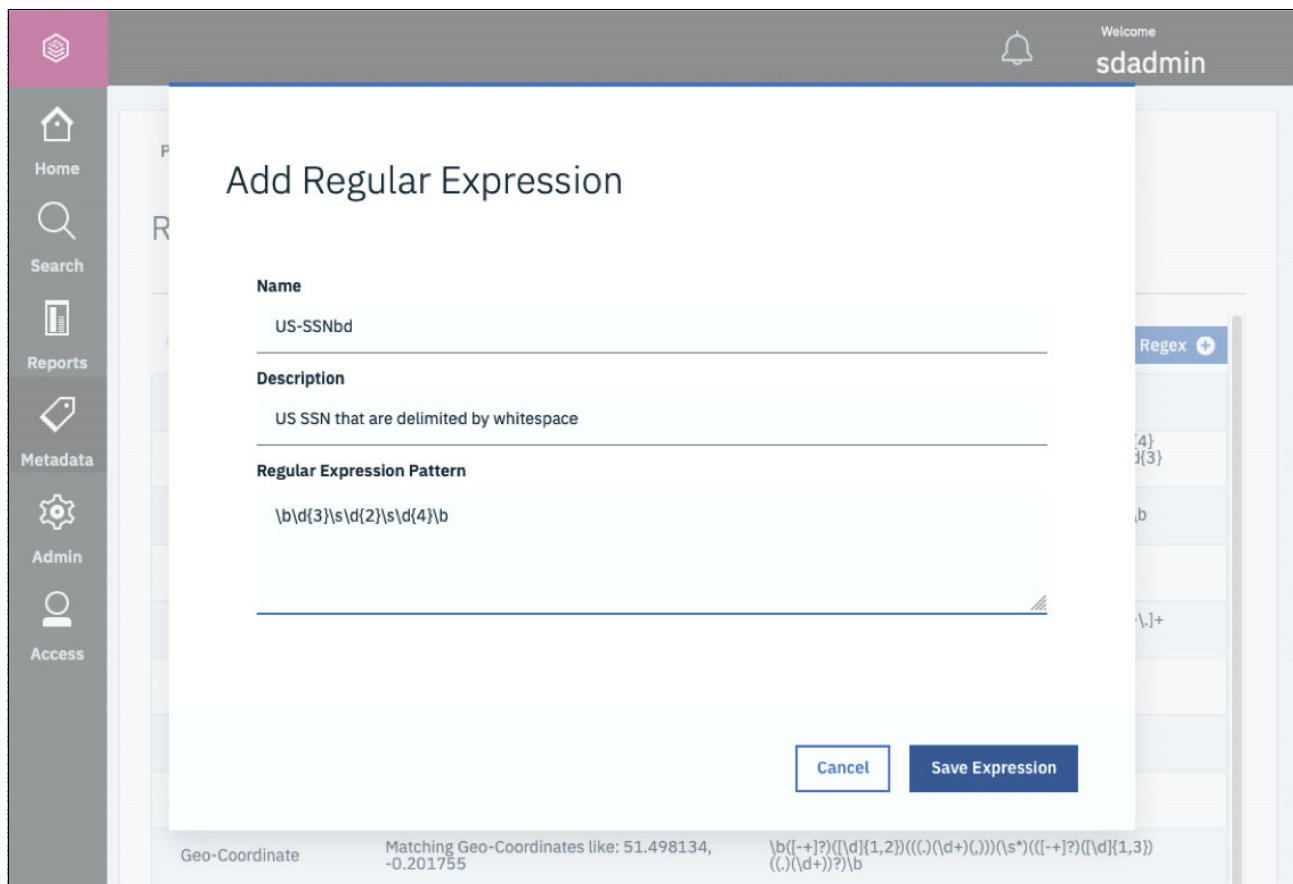


Figure 1-9 Creating a regular expression for use with IBM Spectrum Discover CONTENT SEARCH policies

Defining the CONTENT SEARCH policy

After you verify that the regex is available for your CONTENT SEARCH policy, proceed to defining the policy. As with other policies, browse to the **Metadata** page, select the **Policies** tab, and then select the **Add Policy** option.

Continuing with our example, suppose that we want a policy that identifies an SSN. To define the policy, complete the following steps:

1. Name the policy for example, **identifyFilesWithSSN**.
2. Select the policy type of **CONTENT SEARCH**.
3. Specify a filter pertinent to your scenario.
4. Select the agent that was created for CONTENT SEARCH policies. In this example scenario, that agent is `contentsearchagent`.
5. Identify the tags that the policy is to assign values to, for example, **ssn**.
6. Select the regex that the policy uses when searching the content of your data storage. In this example, we use two types of regex that identify US SSNs.
7. Specify whether you want the policy to set the value of **ssn** to the nine-digit SSN found or if you want to use the tag to identify whether it found an SSN (True) or not (False). In our example, we choose to identify whether the file has an SSN.

After saving the policy configuration, your CONTENT SEARCH policy is ready to use.

A further overview of this capability is included in Chapter 4, “Curating unstructured data for IBM Watson Knowledge Catalog with IBM Spectrum Discover” on page 37 and Chapter 6, “Financial services use case: Personally Identifiable Information detection and data governance” on page 75.

DEEP-INSPECT with Custom Agents

DEEP-INSPECT is a policy that passes lists of files that are based on filter criteria to an analytics agent. It opens the source data file and extracts metadata information from it. The policy passes the data back to IBM Spectrum Discover in the form of tags so that you can do a search and perform the following tasks:

- ▶ Set up a filter to perform a search query that finds the candidates to apply the policy. For example, you can set an action for filtered candidates:
`AUTOTAG, tag1: value, tag2: value`
- ▶ Set a schedule to apply the policy by specifying the following methods:
 - Immediately
 - Periodically

DEEP-INSPECT policies are easily built and run from within the IBM Spectrum Discover GUI, as shown in Figure 1-10.

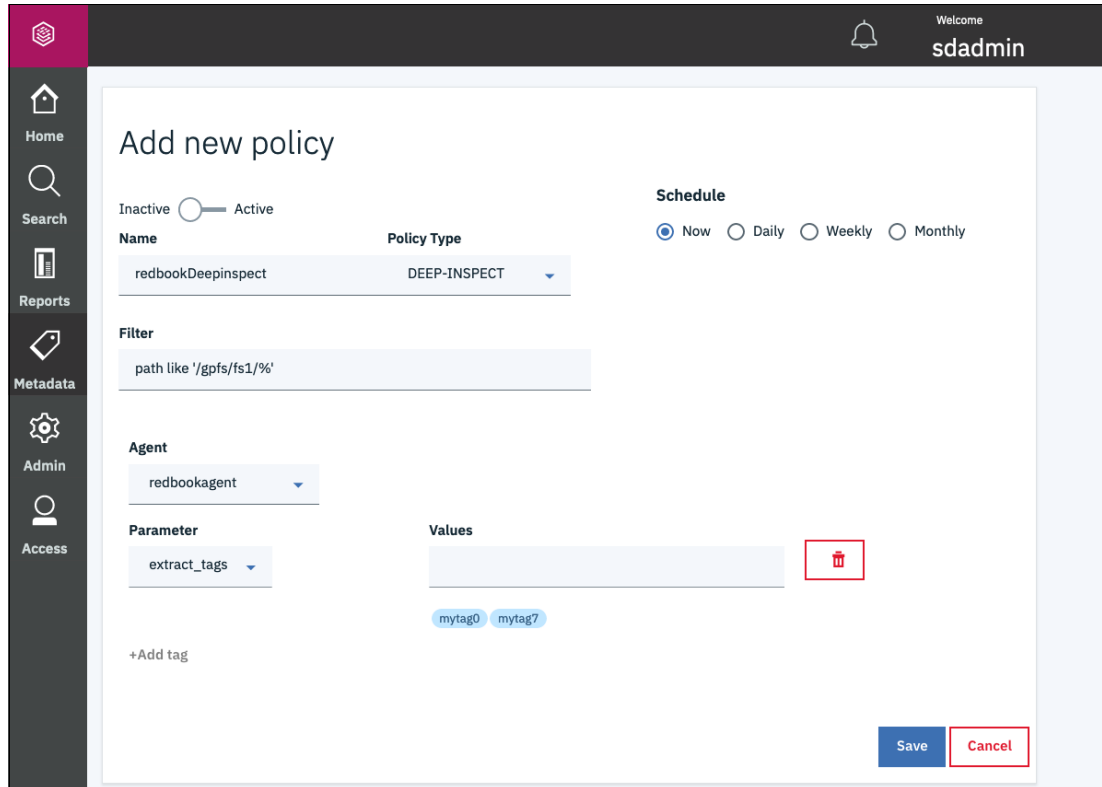


Figure 1-10 DEEP-INSPECT Policy from the IBM Spectrum Discover GUI

In some scenarios, collecting the wanted metadata cannot be accomplished by using AUTOTAG or CONTENTSEARCH policies. Consider a file format that facilitates a specific or proprietary API to collect metadata that is based on that format.

Another example is one in which files include location codes that must be converted to geographical metadata by using a mechanism that performs the conversion of location code to that metadata. To facilitate such use cases (those that do not fit easily into the AUTOTAG or CONTENTSEARCH policy paradigms), IBM Spectrum Discover provides an API to facilitate more complex metadata collection. For more information about API, see [IBM Knowledge Center](#).

To implement a custom metadata collection agent, complete the following steps:

1. Using the IBM Spectrum Discover GUI, define the tag or tags that are associated with the metadata that you plan to collect.
2. Develop an executable (custom agent) by using the REST API that collects the metadata and communicates it to your IBM Spectrum Discover system. This process includes registering your agent, establishing connectivity, and developing the necessary software to communicate with the IBM Spectrum Discover platform.
3. Define your DEEP-INSPECT policy by using the IBM Spectrum Discover GUI.

Defining tags for use with DEEP-INSPECT policies

Defining tags to be used with your DEEP-INSPECT agent and policy is analogous to defining tags for any policy. Because your user-defined agent and policy depend on the tag names, plan and define the tags before moving on to agent development and policy definition.

Developing a DEEP-INSPECT agent to collect metadata

This section describes how to deploy a custom metadata collection agent.

Developing a customer metadata collection agent depends on the IBM Spectrum Discover REST API. For more information about using the API for action agent management, see [IBM Knowledge Center](#).

To operate the DEEP-INSPECT agent by using a IBM Spectrum Discover system, complete the following steps:

1. Obtain the authorization token from the IBM Spectrum Discover system by using the credentials of an IBM Spectrum Discover data admin user.
2. Register your action agent by creating a JSON file by using the details of your action agent.
3. Develop the DEEP-INSPECT agent.

You can use any programming language that you choose to deploy the DEEP-INSPECT agent. At a high level, your agent must establish a connection with the Kafka work and completion topics (message queues). For more information, see [Apache Kafka](#).

4. To create the DEEP-INSPECT policy, complete the following steps:
 - a. Browse to the **Add new policy** pane by selecting the **Metadata** icon in the left pane. Select the **Policies** tab, and then the **Add Policy** option.
 - b. Enter a name for the policy. Select **DEEP-INSPECT** in the Policy Type drop-down list and select your agent from the **Agent** drop-down list.
 - c. Select the **+ Add tag** link, which displays selections for Parameter and Values. In the **Parameter** drop-down list, select **extract_tags**.
 - d. Add each tag that you want the DEEP-INSPECT agent to assign values to the list Values.

An example of defining a DEEP-INSPECT policy that corresponds to our example in this section is shown in Figure 1-11.

Note: For each tag that you want to add to the **Values** list, enter the tag name and then press Return or Enter so that the tag names appear in light blue bubbles below the **Values** entry bar.

The screenshot shows the 'Add new policy' interface in IBM Spectrum Discover. The form is titled 'Add new policy' and includes several sections: 'Inactive' and 'Active' radio buttons, a 'Schedule' section with 'Now', 'Daily', 'Weekly', and 'Monthly' options, a 'Name' field with 'redbookDeepinspect', a 'Policy Type' dropdown menu set to 'DEEP-INSPECT', a 'Filter' field with 'path like '/gdfs/fs1/%'', an 'Agent' dropdown menu set to 'redbookagent', a 'Parameter' dropdown menu set to 'extract_tags', and a 'Values' field containing 'mytag0' and 'mytag7'. There are three red callout boxes with white text: one pointing to the 'Policy Type' dropdown with the text 'Select DEEP-INSPECT policy type', one pointing to the 'Agent' dropdown with the text 'Specify the exact name of the deep inspect agent you registered', and one pointing to the 'Values' field with the text 'Add each tag you want the agent to process'. At the bottom right, there are 'Save' and 'Cancel' buttons.

Figure 1-11 Adding a DEEP-INSPECT policy to an IBM Spectrum Discover system

After the policy is defined, running it sends a list of files or objects to your DEEP-INSPECT agent for processing.

For more information, see 5.1, “Generic healthcare use case” on page 50.

1.5.4 IBM Spectrum Discover Application Catalog and Software Development Kit

IBM Spectrum Discover enables users to customize their metadata extraction capability by using the DEEP-INSPECT policy (see “DEEP-INSPECT with Custom Agents” on page 17), as shown in Figure 1-12 on page 21. That ability is facilitated by the IBM Spectrum Discover Software Development Kit (SDK) to aid in rapid development of agents is provided.

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

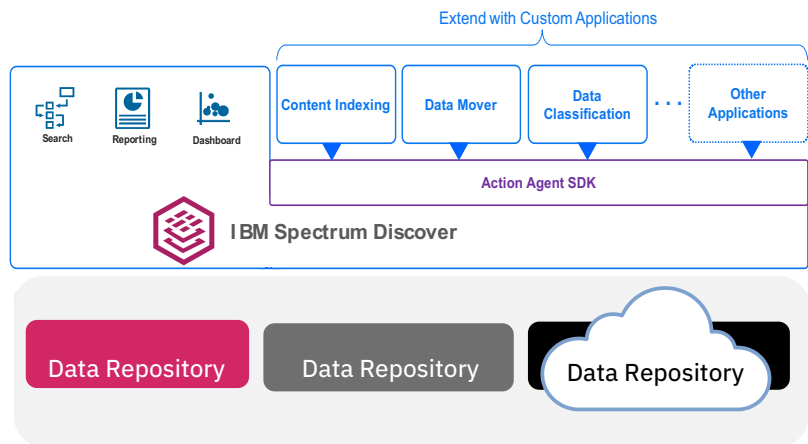


Figure 1-12 IBM Spectrum Discover as an extensible platform for adopting more application tools

The IBM Spectrum Discover SDK is available on [GitHub](#), and a sample application is also available on [Dockerhub](#).

This SDK enables users to use IBM Spectrum Discover as a *platform* that can invoke custom applications and return the values that are identified from that application. This ability enhances the capability and customization options that are available on specific data sets that might be open formats or proprietary data formats.

To expand that capability, IBM established the IBM Spectrum Discover Application catalog as a single place where customers can search, download, and deploy applications for use in IBM Spectrum Discover, as shown in Figure 1-13. The applications in this repository are provided by IBM.

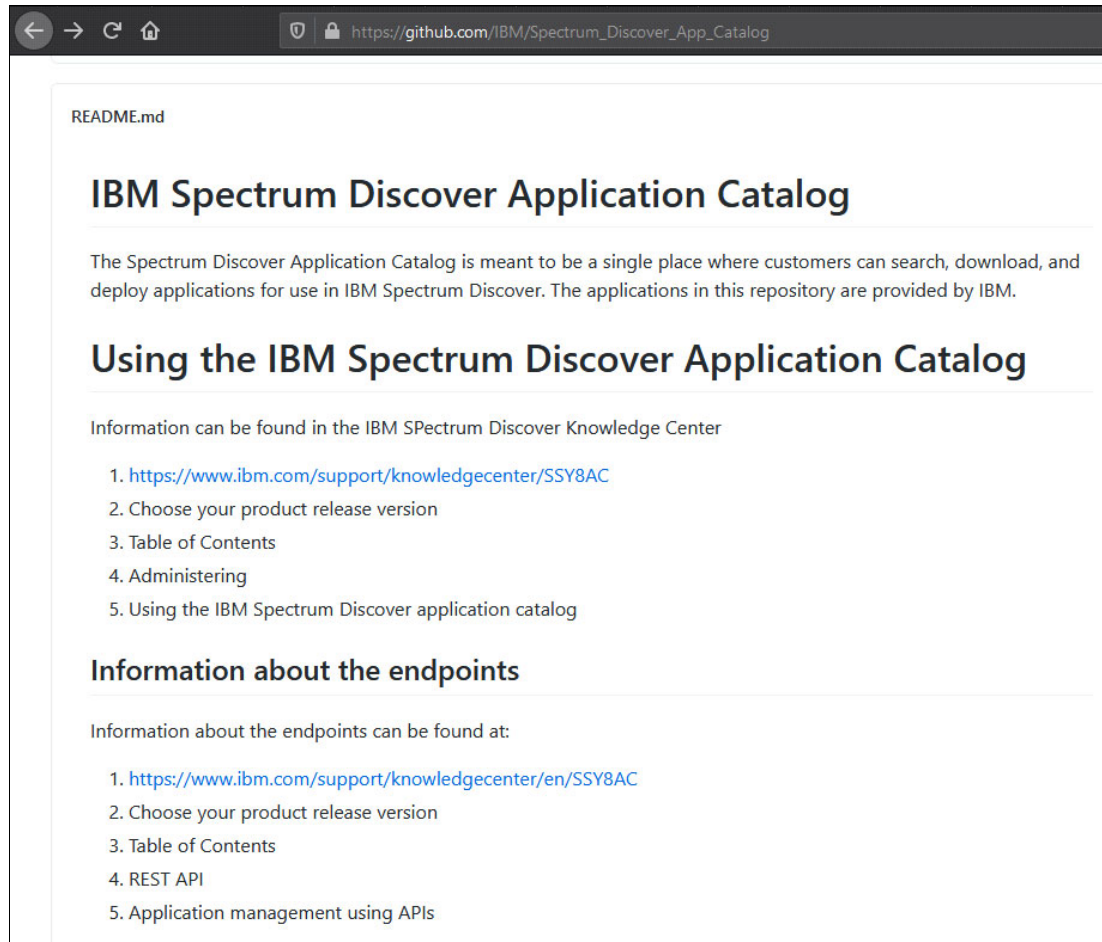


Figure 1-13 GitHub app catalog

For more information about the IBM Spectrum Discover Application Catalog, see [GitHub](#).

1.5.5 Data movement with IBM Spectrum Discover

With Version 2.0.3.1, IBM Spectrum Discover introduced the ability to invoke data movement across select data sources. Now, data administrators or data engineers can move or collect various data sets in various data sources and physical locations into one place based on various characteristics they determined based on the information they see from the tags and reporting tools in IBM Spectrum Discover. Also, it means that those users can invoke a data archiving strategy from view-to-action based on characteristics like data temperature. They identify older data and then invoke the action to move that data between data sources. These approaches are run by IBM Spectrum Discover as a policy action.

There are two different approaches to this data movement:

- ▶ **ScaleILM:** A built-in application to IBM Spectrum Discover that uses the Information Life Cycle Management (ILM) engine that is a part of IBM Spectrum Scale.
- ▶ **External Agent based movement between data sources:** This approach uses an external agent to facilitate that movement.

ScaleILM: IBM Spectrum Scale based

Included with IBM Spectrum Discover V2.0.3 is a built-in application that is named *ScaleILM*, which can be used in data management policies to set up tiers of select set of files or data set on a IBM Spectrum Scale data source connection. The ScaleILM application enables those selected files or data sets to move between internal IBM Spectrum Scale pools in a cluster, for example, between a flash-based pool and an NL-SAS based pool or other configuration, and uses the IBM Spectrum Scale ILM capability. In addition, the ScaleILM tool enables the movement to an external storage repository or archive pools that are managed by IBM Spectrum Archive, which creates an integrated approach to managing the data to the correct tier based on the data sets tags, attributes, and overall information.

External Agent based movement

Included with IBM Spectrum Discover V2.0.3.1 is the capability to move data between data repositories that are being monitored. Using the same type of reporting and understanding of the data that is inherent in IBM Spectrum Discover, data administrators can make decisions and move data based on business requirements. Using the DEEP-INSPECT capability along with validated external applications to run the movement as a policy, such data movement tasks can be orchestrated. As of Version 2.0.3.1, the first external agent application that is validated for this use is [Moonwalk](#).

1.6 Deployment patterns

IBM Spectrum Discover can be deployed by using a single node or multiple nodes, depending on the size of the catalog. For environments with more than 2 billion files to be cataloged, IBM recommends deploying multiple nodes. Consider the following points:

- ▶ Single nodes are for environments in which the file count of the catalog is not greater than 2 billion files.
- ▶ To provide greater performance for environments with more than 2 billion files, and higher availability, deploy three nodes that use shared storage (Red Hat OpenShift deployment only).

There are two different platforms options for the deployment of the IBM Spectrum Discover nodes.

VMware-based deployment

IBM Spectrum Discover is deployed as an Open Virtual Appliance (OVA) image to be deployed on VMware ESXi 6.0 or later.

KVM-based deployment

IBM Spectrum Discover V2.0.3.1 now supports deployment as a Kernel-based Virtual Machine (KVM).

IBM announced a plan to support deployment of IBM Spectrum Discover on Red Hat OpenShift in an upcoming release.



IBM Watson Knowledge Catalog and IBM Cloud Pak for Data overview

This chapter provides an overview of IBM Watson Knowledge Catalog (WKC) and IBM Cloud Pak for Data (IBM CP4D).

This chapter includes the following topics:

- ▶ Overview of Watson Knowledge Catalog
- ▶ Overview of IBM CP4D
- ▶ IBM CP4D

2.1 Overview of Watson Knowledge Catalog

WKC makes high quality, governed, and trusted data accessible for anyone who needs it to make key business decisions. WKC gives you a 360-degree view of your data so that you can easily and quickly discover and uncover insights; index and catalog your data assets; ensure compliance by governing and controlling access to and enforcing policies for governed data assets; and increase collaboration and accountability across teams through a self-service, data-rich experience that auto-guides users to the most relevant data for their task.

Figure 2-1 shows an overview of WKC.

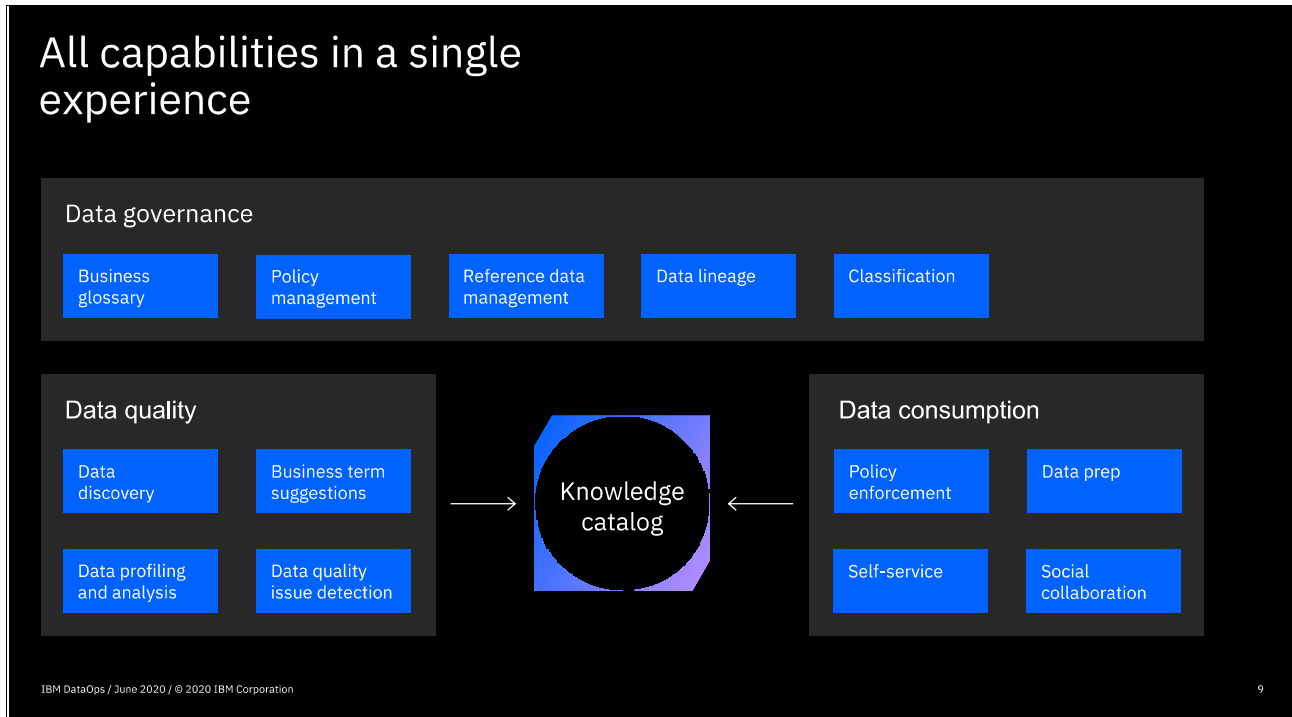


Figure 2-1 WKC capability experience

WKC provides the following capabilities:

- ▶ A single self-service, governed experience to govern (discover, understand) and activate (consume and prepare) your data for artificial intelligence (AI).
- ▶ Identify and inventory where your private, sensitive data is by using 160 predefined or customized data classes.
- ▶ Uncovering data quality issues and auto assignment of business terms in support of data quality initiatives.
- ▶ A single data catalog in a market that is available for deployment on IBM Public Cloud, IBM CP4D, on-premises, Microsoft Azure, and Amazon Web Services (AWS).
- ▶ A unified catalog that is built on top of an open metadata framework to synchronize data catalogs across the customer's business.
- ▶ A persona-specific experience for data quality professionals, data governance teams, data scientists, and business analysts.
- ▶ A Lite version of WKC is embedded in IBM Data and AI offerings (IBM InfoSphere® Information Server, Master Data Management, and IBM Watson Studio), so customers can take advantage of an enterprise data catalog no matter where they are.

WKC is available as several packaging options:

IBM CP4D

For enterprise customers who want IBM CP4D with comprehensive catalog, policy enforcement, data quality, data governance, and self-service capabilities.

IBM Watson Data Platform on IBM Cloud

For customers looking for a software as a service (SaaS) offering. Watson Data Platform is part of IBM Cloud, and includes WKC, Watson Studio, Watson Machine Learning, and Open Scale. Customers can try WKC for free through its Lite edition or purchase subscriptions of Standard or Professional.

2.2 Overview of IBM CP4D

AI and big data are no longer future technologies. Using AI, big data, and analytics is imperative in today's business environment. When you need the best solutions for your big data and AI workloads, the right choice is IBM. IBM CP4D is a fully integrated data and AI platform that modernizes how businesses collect, organize, and analyze data to infuse AI throughout their organizations. It is unified, modular, and deployable anywhere.

Figure 2-2 shows the four pillars of action.

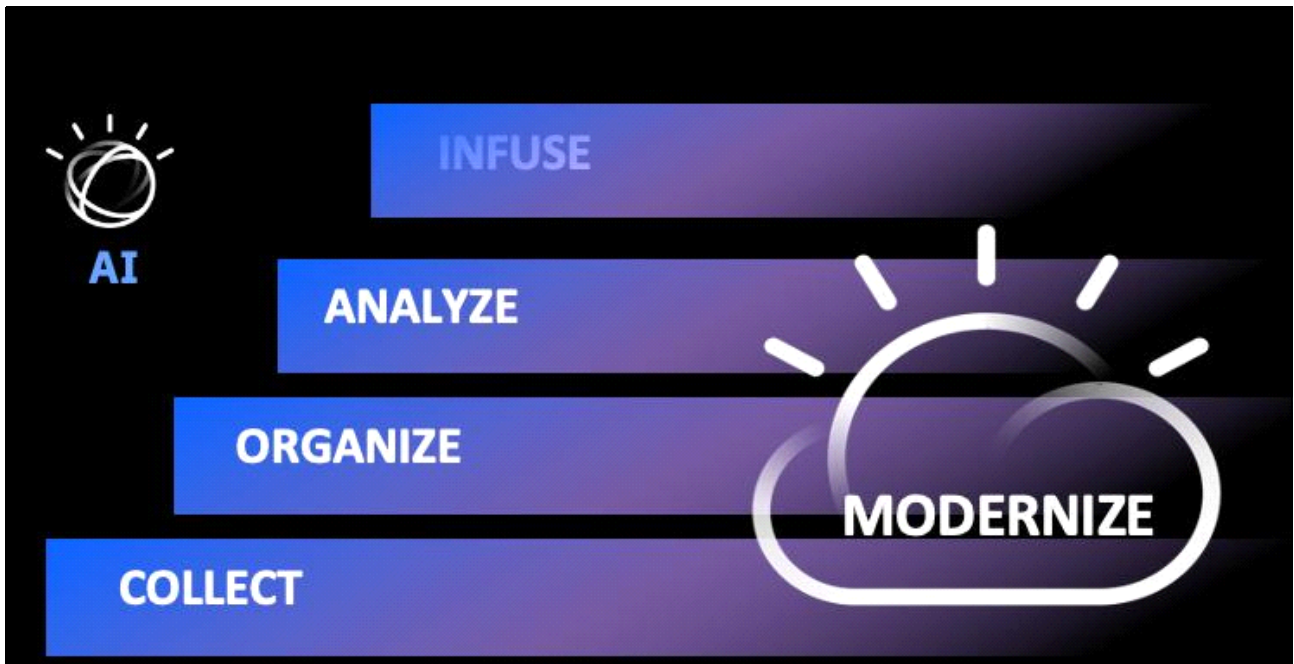


Figure 2-2 The four pillars of action of IBM CP4D

2.2.1 IBM CP4D and WKC

WKC helps organizations deliver trusted and meaningful data by providing a secure, enterprise catalog management platform that is supported by a data governance framework. A catalog stores data and knowledge, and the data governance framework ensures that data access and data quality comply with required business rules and standards.

WKC in IBM CP4D role-based permissions supports roles from data engineers integrating data, data governance teams working on stewardship and data quality, data citizens who might be business users analyzing data, and data scientists working on data science projects.

Data stewards and data quality analysts govern and curate data to provide business-ready data assets that are easy to find and secure.

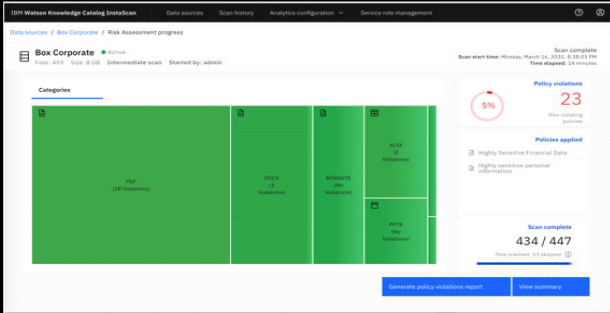
Curation includes discovering, classifying, and understanding all types of data to ensure that the data is complete, applicable, and meets quality standards.

WKC key features are:

- ▶ Real-time data virtualization
- ▶ Automated data discovery and metadata generation
- ▶ A machine learning (ML) extracted business glossary from most common regulatory terms
- ▶ Dynamic data masking to protect sensitive data
- ▶ Automated scanning and risk assessment of unstructured data by using Watson Catalog InstaScan (shown in Figure 2-3)

Watson Knowledge Catalog InstaScan

Mitigate risks in a fraction of the time



Business Challenges

Need to quickly establish sustainable data privacy processes for regulations like GDPR and CCPA

Technical impossible to keep up with growth of unstructured data in cloud and running risk assessments over the wire

Business Value

Conduct unstructured risk assessment in a *fraction of the time* using industry recognized sampling algorithms

Ongoing compliance check to gain confidence and *share with regulators* – get clean, stay clean

InstaScan on Cloud Pak for Data

- Unstructured risk assessment and compliance check
- Quickly determine which areas have high concentration of sensitive information and prioritize hot spots
- Automatically apply classification labels to data that violates corporate policies
- Create and share ongoing compliance reports with CISO or regulators

Figure 2-3 WKC InstaScan

2.3 IBM CP4D

One of the most important features for every company is how fast and efficiently they can extract value from their data. IBM CP4D speed time to value with a single platform that integrates data management, data governance, and analysis, which bring greater efficiency.

Figure 2-4 on page 29 shows the capabilities, enhancements, and business impact of IBM CP4D.

28 Cataloging Unstructured Data in IBM Watson Knowledge Catalog with IBM Spectrum Discover


Cloud Pak for Data v3.0

New Capabilities & Enhancements		Business Impact
Modernize: Data Modernization	Unified user experience	Accelerate time-to-value with agile workflows and execution processes across data, services, and user roles
	OpenShift v4.3 and v3.11 support	Reduce costs with 66% faster development lifecycle and migrate security risks with certified security assurance
	OpenShift Container Storage (OCS) support	Reduce costs of data footprint with software-defined low-cost storage
Collect: Data Modernization	Data access enhancements – including tighter governance integration with Data Virtualization and onboarding Db2 Big SQL	Reduce costs of data architecture with support for object stores and separation of compute and storage
Organize: DataOps	Watson Knowledge Catalog – new InstaScan capability	Mitigate risks with unstructured risk assessments in a fraction of the time
Analyze: AI Lifecycle Automation	AutoAI enhancements – now with zero lock-in	Accelerate time-to-value by building models in minutes instead of days or weeks, with zero lock-in
Infuse: AI for Financial Operations	Planning Analytics	Accelerate time-to-value with 80% faster planning system processing

Figure 2-4 IBM CP4D

Here are the benefits of IBM CP4D:

- ▶ Eliminates data silos and connects all your data.
- ▶ Automates and governs the data and AI lifecycle.
- ▶ Extends your platform with extra services.
- ▶ Is a single unified platform.
- ▶ Has best-in-class data virtualization.
- ▶ Uses built-in data governance
- ▶ Has extensible APIUs and ecosystem.
- ▶ Has unified data and AI services.
- ▶ Unlocks IT savings.
- ▶ Increases enterprise-level governance and security.
- ▶ Consolidates tools.
- ▶ Reduces storage costs.
- ▶ Minimizes data movement expenses.
- ▶ Improves infrastructure savings.
- ▶ Fully supports multicloud environments like AWS, Azure, Google Cloud, IBM Cloud, and private cloud deployment.



IBM Spectrum Discover integration with IBM Watson Knowledge Catalog architecture and benefits

This chapter describes the architecture and benefits of IBM Spectrum Discover with Watson Knowledge Catalog (WKC), and describes how to configure IBM Spectrum Discover to connect to WKC and export curated assets.

This chapter includes the following topics:

- ▶ Solution architecture
- ▶ Connecting IBM Spectrum Discover to Watson Knowledge Catalog
- ▶ Exporting assets from IBM Spectrum Discover to Watson Knowledge Catalog
- ▶ Using assets in Watson Knowledge Catalog

3.1 Solution architecture

IBM Spectrum Discover curates and organizes vast amounts of unstructured data from storage sources on premises and in the cloud, and creates an enriched, structured view of the content that is maintained in the IBM Spectrum Discover catalog. Users use the IBM Spectrum Discover GUI or REST application programming interface (API) to select assets from the IBM Spectrum Discover catalog and publish them into one or more data catalogs in WKC. Custom tags and enrichments that are associated with the assets from IBM Spectrum Discover may also be published into WKC. After the assets are registered in WKC, users can create projects and collaborate with peers on data science projects, establish data governance policies, and use the suite of tools and capabilities in IBM Cloud Pak for Data (IBM CP4D) on the assets.

Unstructured data assets that are registered in WKC by IBM Spectrum Discover can be combined with more WKC assets, such as structured data, semi-structured data, and other cloud based unstructured data, such as assets from Box to provide an end to end view of all data across an enterprise.

Figure 3-1 illustrates this integration.

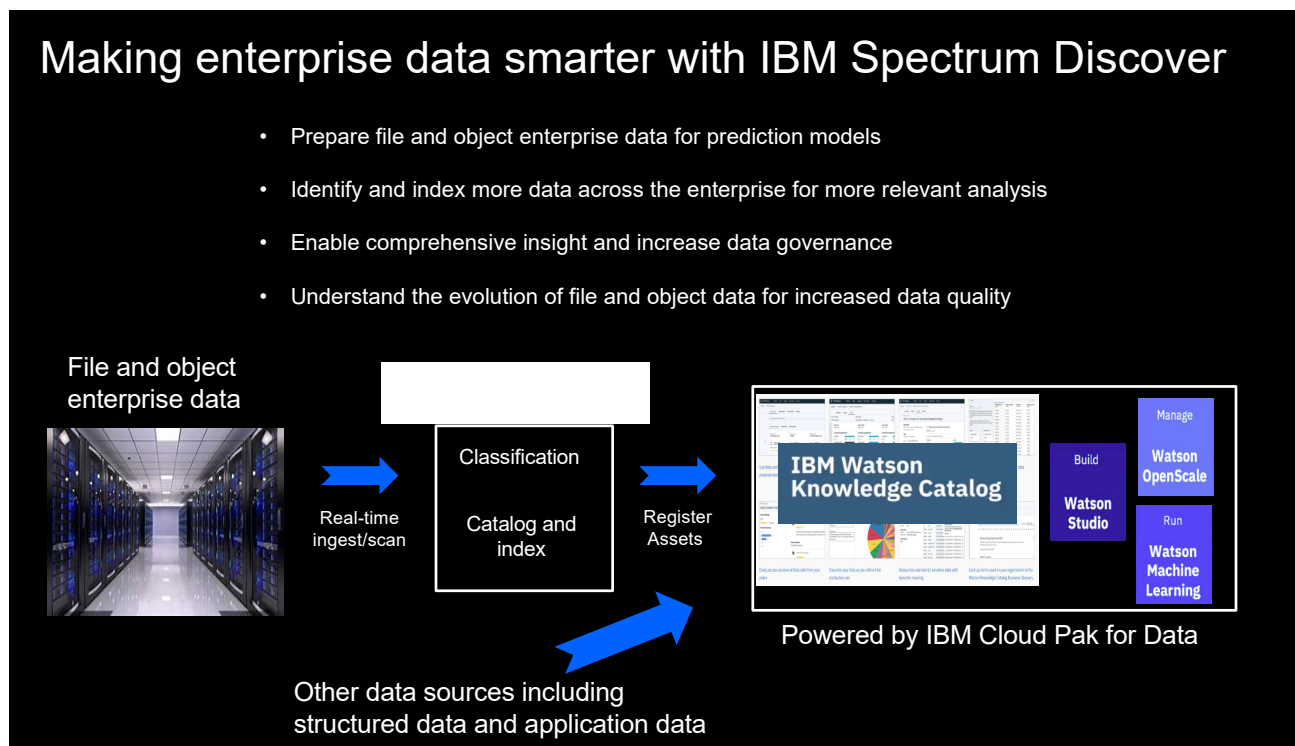


Figure 3-1 IBM Spectrum Discover and Watson Knowledge Catalog integration

3.1.1 Asset registration process

WKC uses catalogs to organize resources for projects. The resources include data assets, analytical assets, connections, and collaborators who can use the assets. The assets in a catalog can be enriched with other governance artifacts and governed by policies to enforce data protection rules.

When IBM Spectrum Discover exports assets to a catalog in WKC, an asset metadata record is created in the user-specified catalog, and an attachment or link to the asset location is created. The asset attachment in WKC must be addressable through a URL from a Simple Storage Service (S3) based object storage instance. If the asset to be exported is accessible through an S3 endpoint such as Ceph S3, IBM Cloud Object Storage (IBM COS), or the IBM Spectrum Scale S3 object interface, the connection mapping in IBM Spectrum Discover (WKC_CONNECTION_MAP) can be configured, and IBM Spectrum Discover uses the existing URL from the S3 endpoint when attaching the asset to the metadata record in WKC.

If the asset to be registered into WKC is in a storage system that is not accessible through an S3 endpoint, the asset is copied into a customer-provided S3 bucket that is attached to WKC, and the URL of the asset in the S3 bucket is used as the attachment link. More specifically, data in the following storage platform and protocol combinations is automatically copied to a user-provided S3 bucket when they are exported by IBM Spectrum Discover into WKC:

- ▶ IBM Spectrum Scale POSIX/Network File System (NFS) / Server Message Block (SMB) / HDFS
- ▶ Dell/EMC Isilon NFS / SMB
- ▶ NetApp NFS / SMB

Note: Registering assets from IBM Spectrum Protect data sources by IBM Spectrum Discover automatically into WKC is not supported currently.

3.2 Connecting IBM Spectrum Discover to Watson Knowledge Catalog

IBM Spectrum Discover uses IBM Watson REST APIs to export assets into WKC. The WKCCConnector App is included with the standard installation, but requires a one-time setup. After the WKCCConnector App is set up correctly in IBM Spectrum Discover, the “Export Data” button is enabled as an option in the UI for Search result. For IBM Spectrum Discover V2.0.3.1, the WKCCConnector App can be deployed for one WKC instance with multiple catalogs.

The one-time configuration process entails configuring IBM Spectrum Discover with the REST API endpoint and the IBM Watson API credentials to establish a connection to WKC with an existing catalog. An S3 asset connection in WKC must also be created. To enable asset registration that uses the original data in place and does not copy it to a new S3 endpoint, the S3 asset connection that is created in WKC must be the same S3 endpoint that is registered in IBM Spectrum Discover.

For more information about configuring the connection to WKC, see [IBM Knowledge Center](#).

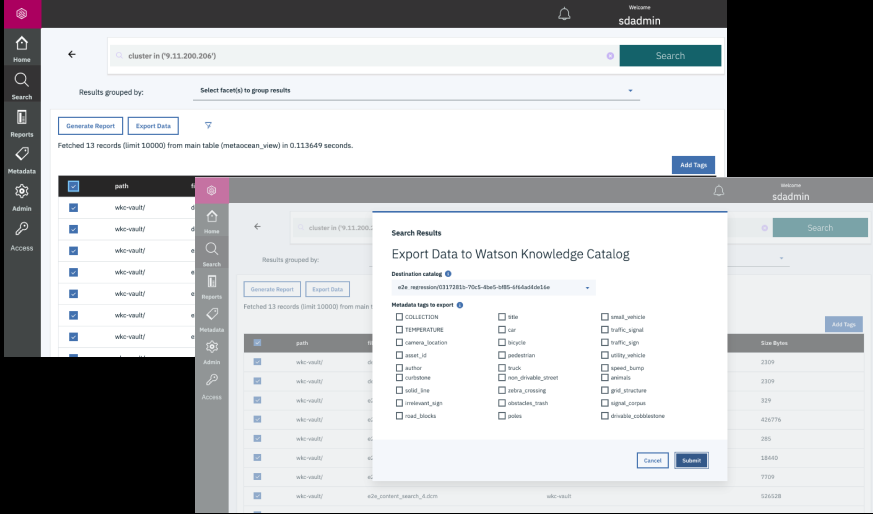
3.3 Exporting assets from IBM Spectrum Discover to Watson Knowledge Catalog

Users use the IBM Spectrum Discover search interface to find assets of interest to be exported to WKC in IBM CP4D. After the data set is identified, users select the assets and initiate a workflow to automatically export the assets to WKC. When exporting assets to WKC, the file / object name is exported by default, and the user can specify the additional metadata to be registered by including one or more custom tags.

Figure 3-2 illustrates this workflow.

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



The screenshot shows the IBM Spectrum Discover interface. At the top, there's a search bar with 'cluster in (*9.11.200.20k)' and a 'Search' button. Below it, there are buttons for 'Generate Report' and 'Export Data'. A message indicates 'Fetched 13 records (limit 10000) from main table (metaaccan_view) in 0.113649 seconds.' A table of results is visible, with columns for 'path', 'who-vault', and 'Size Bytes'. A modal window titled 'Export Data to Watson Knowledge Catalog' is open, showing a 'Destination catalog' dropdown and a grid of 'Metadata tags to export' with checkboxes. The modal also has 'Cancel' and 'Submit' buttons.

Figure 3-2 Exporting data sets from IBM Spectrum Discover to Watson Knowledge Catalog

3.3.1 IBM Spectrum Discover tag to WKC tag mapping

A tag in WKC consists of one string. It can contain spaces, letters, numbers, underscores, dashes, and the symbols # and @. The tags can be assigned to an artifact and used as metadata to simplify searching for governance in WKC.

IBM Spectrum Discover implements a key-value type tagging implementation, and WKC implements a value-type tagging implementation. When exporting tags from IBM Spectrum Discover to WKC, the original tags are converted from a key-value implementation to a value-type implementation that uses a single value. For example, in IBM Spectrum Discover, if a tag is defined as Author and the value is Mark Twain (Samuel Clemens), in IBM WKC the tag becomes author: Mark Twain (Samuel Clemens) as a single value. Figure 3-3 on page 35 illustrates the IBM Spectrum Discover Author tag key and value.

After the data is exported from IBM Spectrum Discover to WKC, the details of the data assets, including the tags, can be viewed in the catalog, and the content of the data can be viewed in WKC.

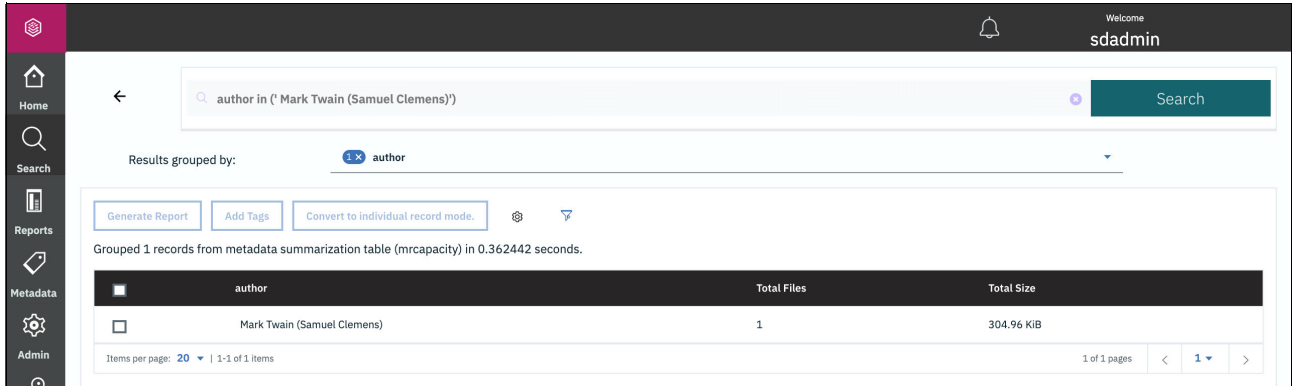


Figure 3-3 IBM Spectrum Discover tag key value example

Figure 3-4 illustrates the equivalent tag value mapping in WKC.

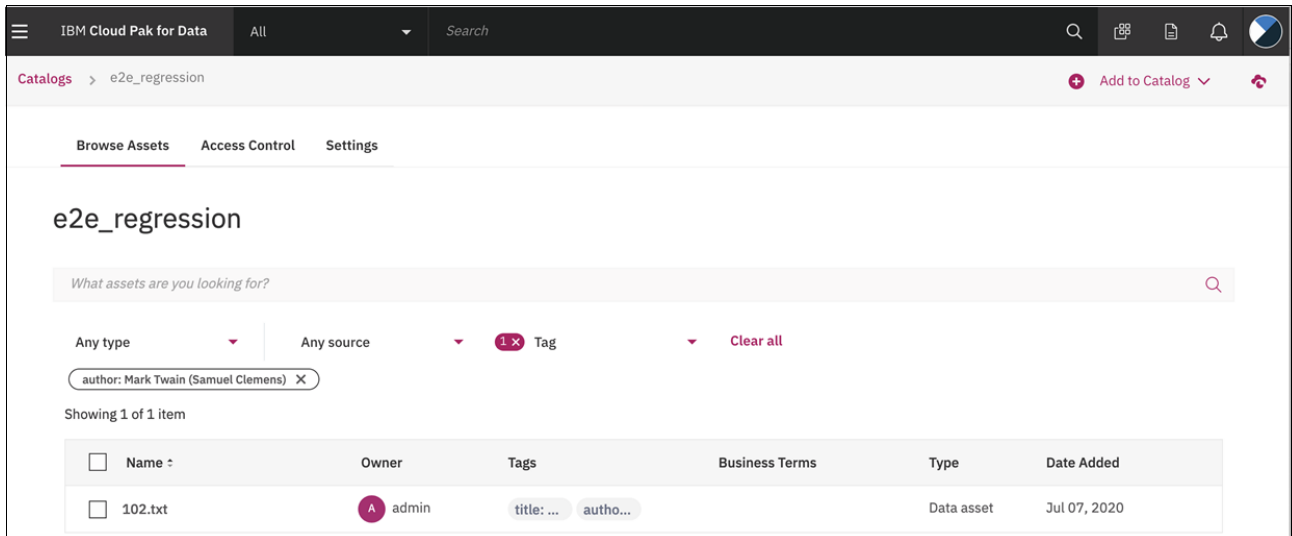


Figure 3-4 Watson Knowledge Catalog tag value conversion example

3.4 Using assets in Watson Knowledge Catalog

After assets from IBM Spectrum Discover are successfully registered in WKC, users can access the assets in the catalog, create projects, use the assets with other tools in IBM CP4D, and create and enforce data governance policies that are based on tags. Detailed examples of these capabilities are described in 4.2, “Using assets in IBM CP4D and Watson Knowledge Catalog” on page 43.



Curating unstructured data for IBM Watson Knowledge Catalog with IBM Spectrum Discover

This chapter provides an example that uses IBM Spectrum Discover to curate and organize unstructured text in IBM Spectrum Scale by searching and extracting keywords from the text. The data along with the insights from the data that is gathered by IBM Spectrum Discover is automatically made available in IBM Watson Knowledge Catalog (WKC) and IBM Cloud Pak for Data (IBM CP4D), where users can create projects and collaborate with peers on data science projects, establish data governance policies, and use the suite of tools and capabilities in IBM CP4D on the assets.

This chapter includes the following topics:

- ▶ Data curation workflow
- ▶ Creating tags in IBM Spectrum Discover
- ▶ Creating regular expressions
- ▶ Creating a content inspection policy
- ▶ Searching by title and author
- ▶ Using assets in IBM CP4D and Watson Knowledge Catalog

4.1 Data curation workflow

The data set that is used in this example is a collection of electronic books on an IBM Spectrum Scale file system from Project Gutenberg. The books in the collection are mostly older literary works that were published before 1924 and for which U.S copyright expired. An IBM Spectrum Scale file system was scanned by IBM Spectrum Discover to catalog the system metadata that is associated with the Gutenberg data set.

A content inspection policy in IBM Spectrum Discover extracts the titles and authors of the books and indexes these values as custom tags in the IBM Spectrum Discover database, enabling users to quickly and easily search the data set by title and author. Documents that were authored by Charles Dickens are selected and automatically registered into WKC in IBM CP4D. The IBM Spectrum Scale unified file and object feature is used to present the documents through a Simple Storage Service (S3) interface to WKC.

Figure 4-1 illustrates this workflow.

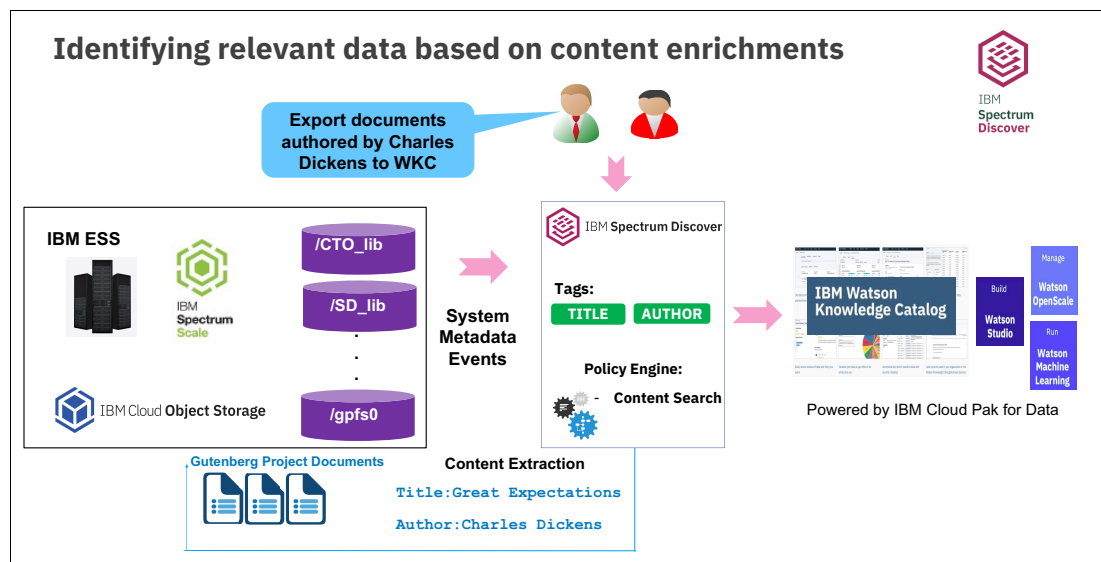


Figure 4-1 Exporting documents that were authored by Charles Dickens to WKC

4.1.1 Creating tags in IBM Spectrum Discover

To create an author tag in IBM Spectrum Discover, complete the following steps:

1. Select **Metadata** → **Tags**, and click **Add tag**.
2. Enter the name of the tag, such as "author", click **Open tag type**, and click **Submit**.

Figure 4-2 on page 39 shows the tag creation dialog.

New Organizational Tags

Name
author

Type
Open

Values
Press 'Enter' key to add the tag to the list
Add a value

Cancel Submit

Figure 4-2 Tag creation dialog

Then, create a title tag in IBM Spectrum Discover by completing the following steps:

1. Select **Metadata** → **Tags**, and click **Add tag**.
2. Enter the name of the tag, such as “title”, select the **Characteristics tag** type, and click **Submit**.

4.1.2 Creating regular expressions

To create a regular expression (regex) to search for the author in the Gutenberg ebook data set, complete the following steps:

1. Select **Metadata** → **Regular Expression**, and select **Add Regex**.
2. Enter a name for the regex, such as “Gutenberg-author”.
3. Enter a description for the regex, such as “Author of Gutenberg Collection”.
4. Enter the regex pattern, for example, Author:(.*)\$.

Figure 4-3 shows the Add Regular Expression window.

The screenshot shows a window titled "Add Regular Expression". It contains three input fields:

- Name:** Gutenberg-author
- Description:** Author of Gutenberg Collection
- Regular Expression Pattern:** Author:(.*)\$

Figure 4-3 Add Regular Expression for author

To create a regex to search for the title in the Gutenberg ebook data set, complete the following steps:

1. Select **Metadata** → **Regular Expression**, and select **Add Regex**.
2. Enter a name for the regex, such as “Gutenberg-title”.
3. Enter a description for the regex such as “Title of Gutenberg Collection”.
4. Enter the regex pattern, for example, Title:(.*)\$

4.1.3 Creating a content inspection policy

To create a content inspection policy to automatically extract the title and author from the Gutenberg data set, complete the following steps:

1. Select **Metadata** → **Policies**, and select **Add Policy**.
2. Make the policy active by clicking the slider.
3. Provide a name for the policy, for example, “gutenberg_title_author”.
4. Select **content-search** as the policy type.
5. Provide the filtering criteria that specifies which documents are inspected. In this example, “path like '/mnt/datadump/gutenberg%' and filetype='txt'” is specified to indicate that text documents in the /mnt/datadump/Gutenberg directory are inspected.
6. Select **contentsearchagent** as the application that performs the inspection.
7. Click **Add row**, select **author** as the tag, select **Gutenberg-author** as the regex, and select **Value matching expression**.
8. Click **add row**, select **title** as the tag, select **Gutenberg-title** as the regex, and select **Value matching expression**.
9. Click **Save**.

Figure 4-4 on page 41 shows the Add new policy window.

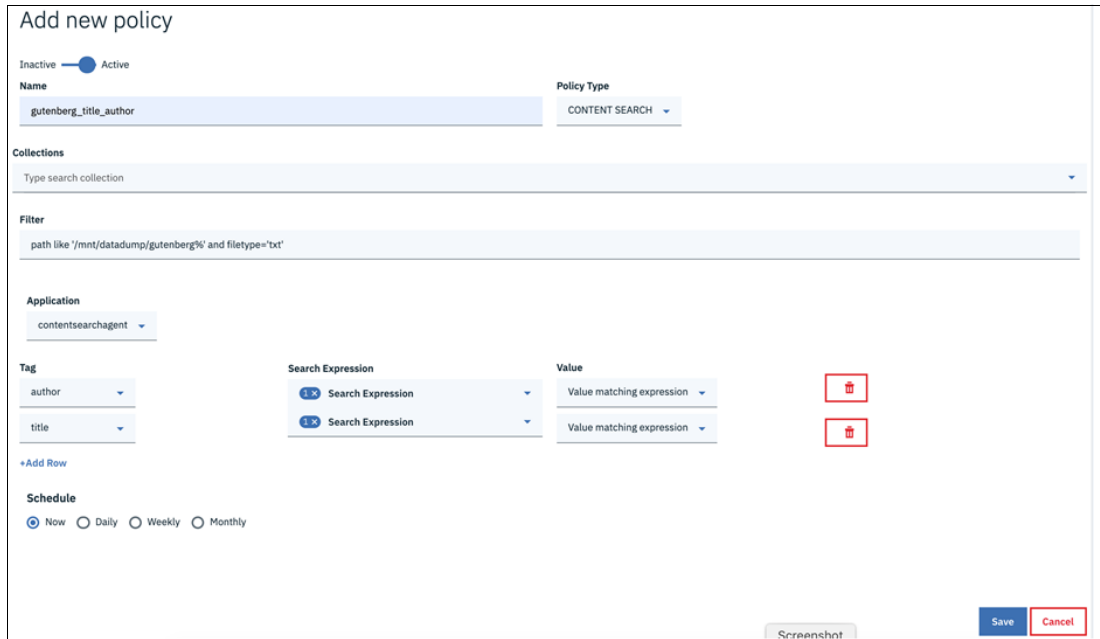


Figure 4-4 Add new policy window

4.1.4 Searching by title and author

To search by title and author, complete the following steps:

1. Click **Search**, and select **author** as the tag for the visual search criteria.
2. Type Charles_Dickens into the visual search bar and select all, as shown in Figure 4-5.

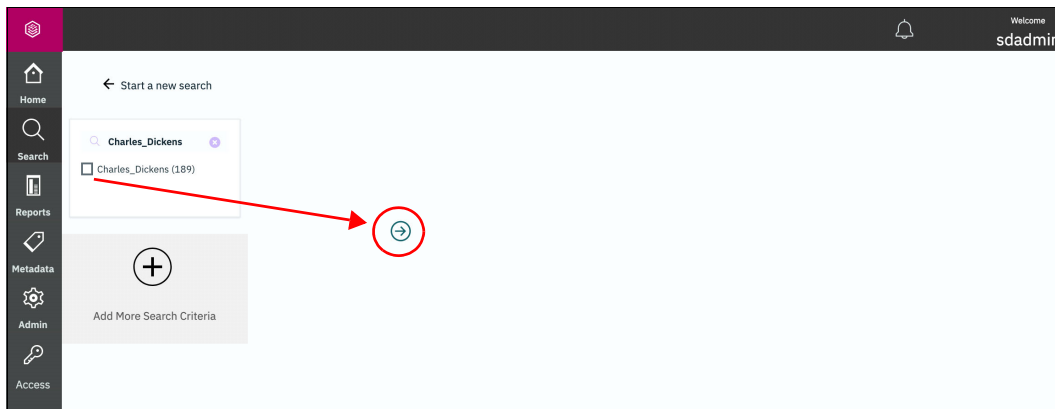


Figure 4-5 Search by Author

3. Click the arrow on the right.

- Select the results that are grouped by the author Charles_Dickens and click **Convert to Individual Record Mode**, as illustrated in Figure 4-6.

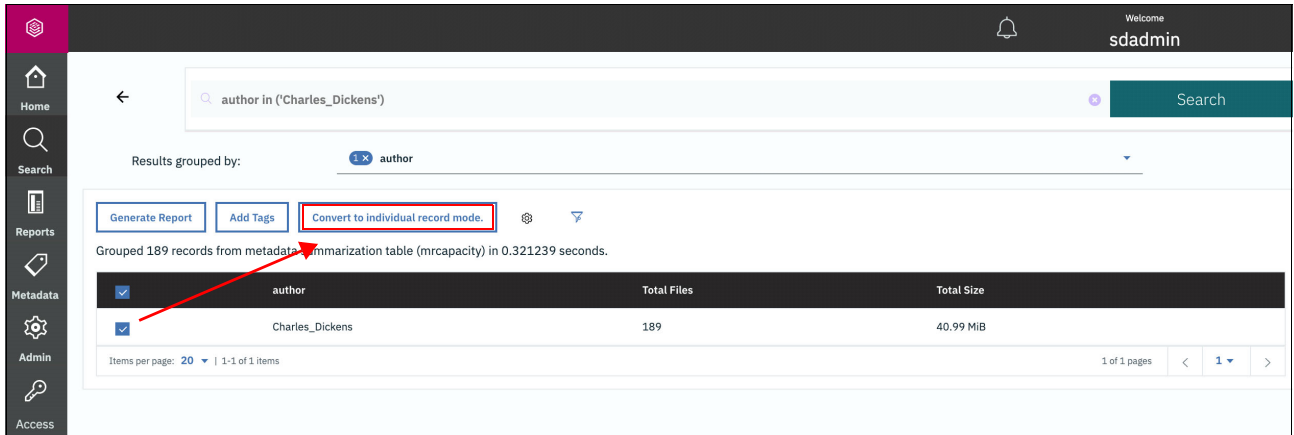


Figure 4-6 Convert to individual record mode selection

- Select all the individual records, and select **Export Data** to export to WKC, as illustrated in Figure 4-7.

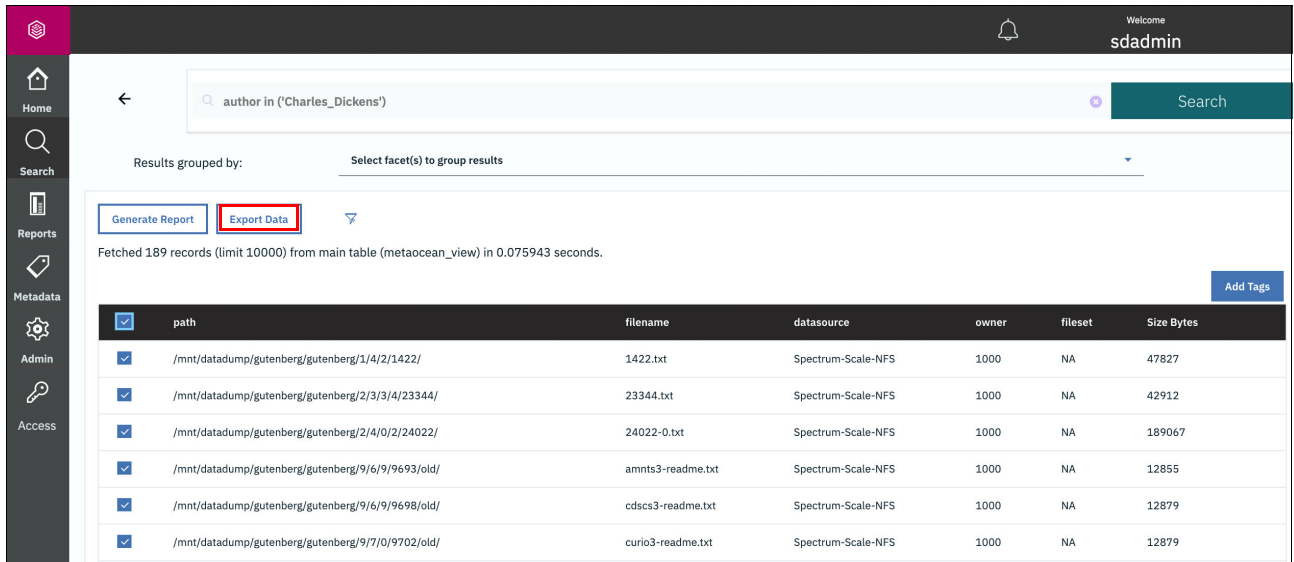


Figure 4-7 Selecting data to export data to Watson Knowledge Catalog

- In the Export to Watson Knowledge Catalog dialog (Figure 4-8 on page 43), select the **WKC catalog** and specify the **title** and **author** tags to persist them to the catalog.

Search Results

Export Data to Watson Knowledge Catalog

Destination catalog ⓘ

c2d9b102-0b82-41eb-b239-dc8eef309a78 ▾

Metadata tags to export ⓘ

- COLLECTION
- title
- TEMPERATURE
- asset_id
- camera_location
- author

Cancel Submit

Figure 4-8 Export Data to Watson Knowledge Catalog dialog

The assets are automatically exported to the specified WKC instance.

4.2 Using assets in IBM CP4D and Watson Knowledge Catalog

This section provides examples of using unstructured assets that are exported by IBM Spectrum Discover with WKC and IBM CP4D.

4.2.1 Browsing and managing assets in a catalog

To browse and manage assets in a catalog, complete the following steps:

1. Go to the catalog in WKC that contains the assets that are exported from IBM Spectrum Discover. Select the **author:Charles_Dickens** tag, as shown in Figure 4-9.

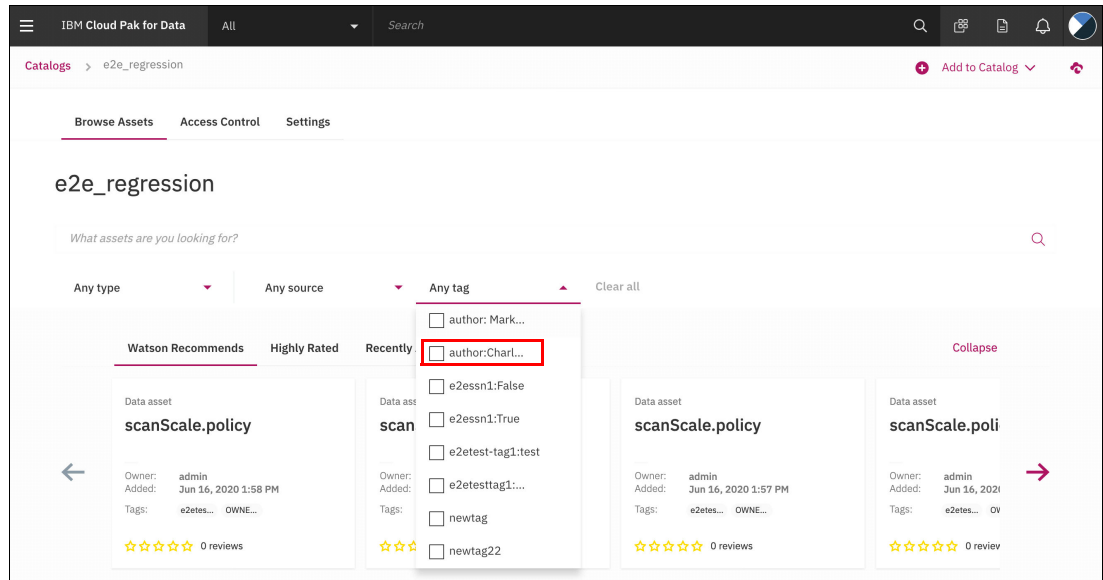


Figure 4-9 Filter by tag in Watson Knowledge Catalog

Figure 4-10 provides an example of the results that are filtered by the **author:Charles_Dickens** tag.

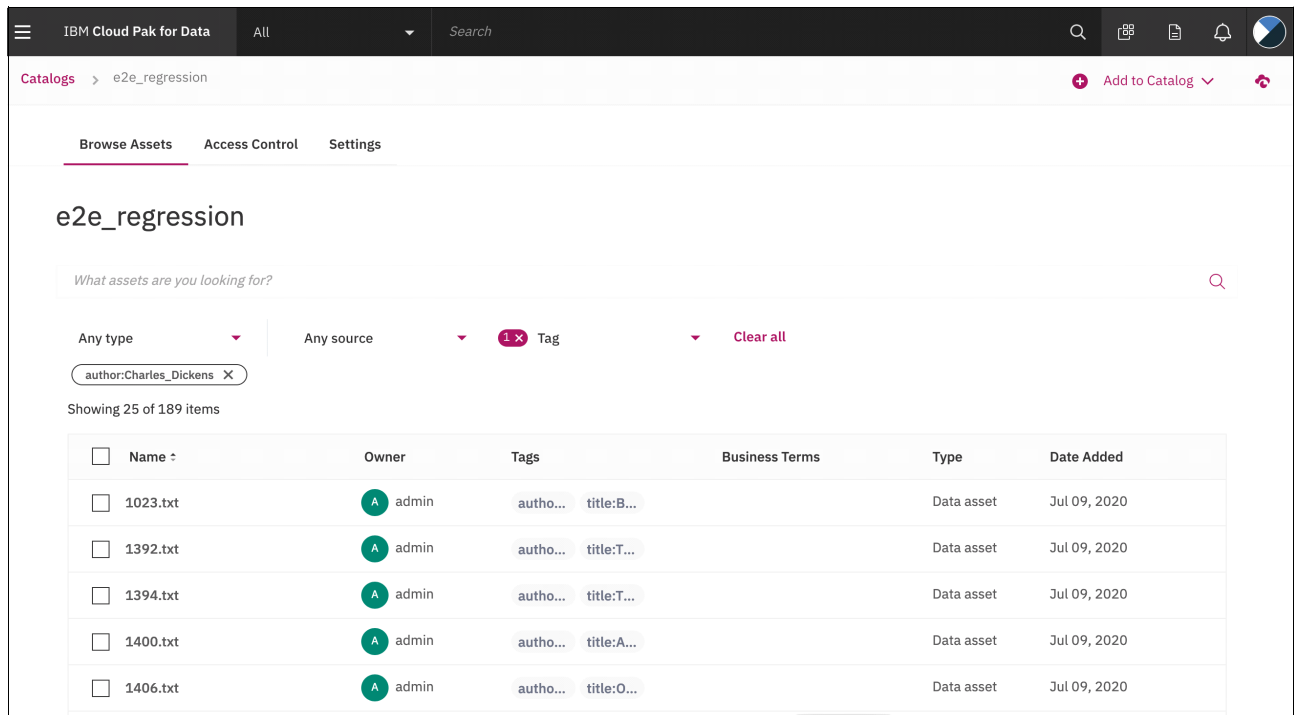


Figure 4-10 Filtered search by tag result in Watson Knowledge Catalog

2. Click an asset to preview it. Both the title and author tags that were created and extracted from the document by IBM Spectrum Discover are shown as tags, as indicated in Figure 4-11.

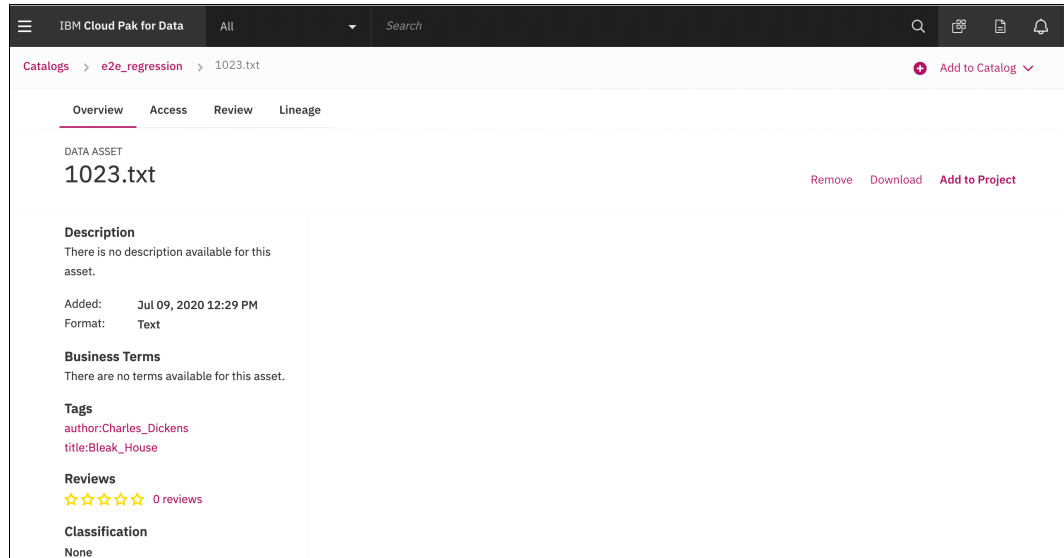


Figure 4-11 Asset preview in Watson Knowledge Catalog

4.2.2 Creating projects from assets in Watson Knowledge Catalog

Assets in WKC can be used in projects in IBM CP4D. To do so, complete the following steps:

1. Click the horizontal bars in the upper left of the IBM CP4D GUI and select **Projects**.
2. Click **New Project** at the upper right of the Projects page, as indicated in Figure 4-12.

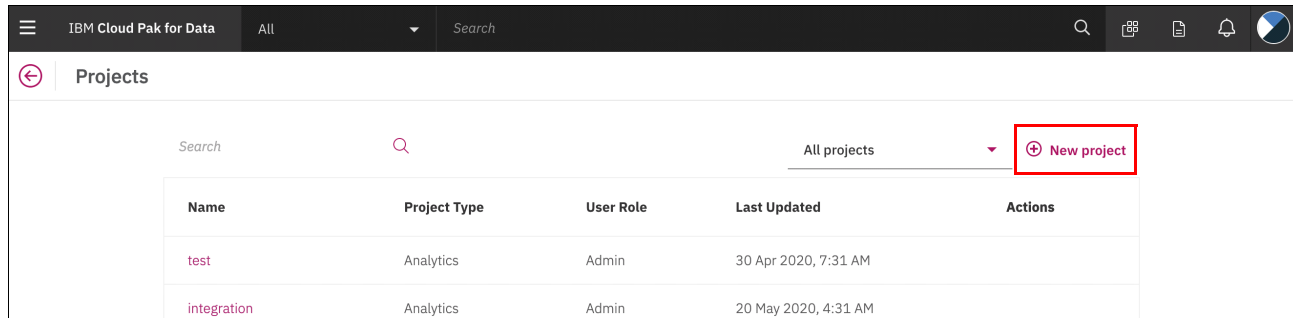


Figure 4-12 IBM CP4D Projects page

3. Select **Analytics Project**, and click **OK**.

4. A window opens where you can create a project, as shown in Figure 4-13. Select **Create an empty project**.

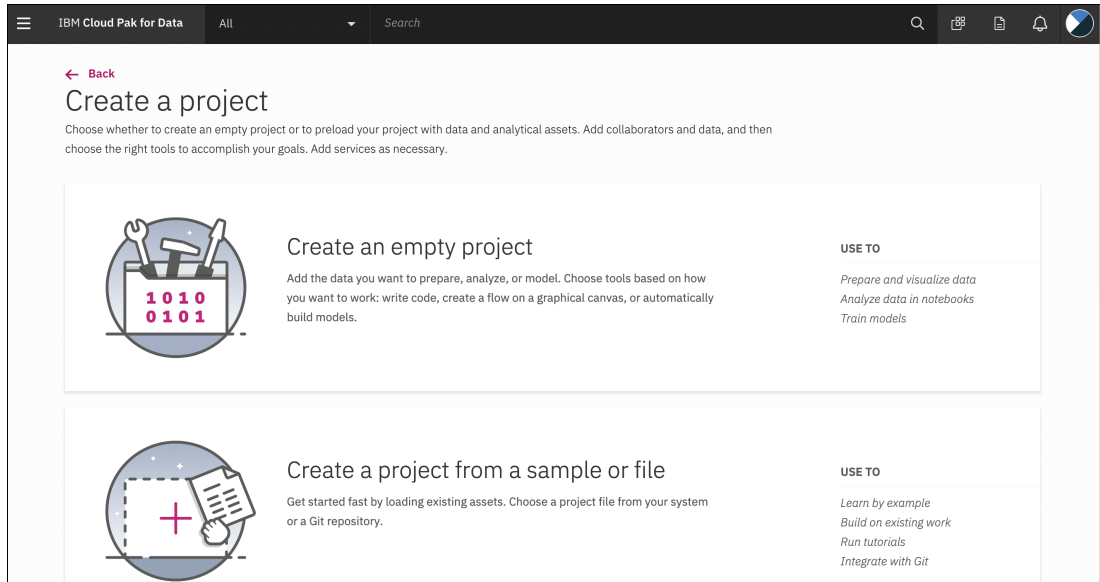


Figure 4-13 Creating a project in Watson Knowledge Catalog

5. Provide a **name** and **description** for the project, optionally select the **Integrate project with Git** option, and click **Create**. Figure 4-14 shows an example of the project workspace.

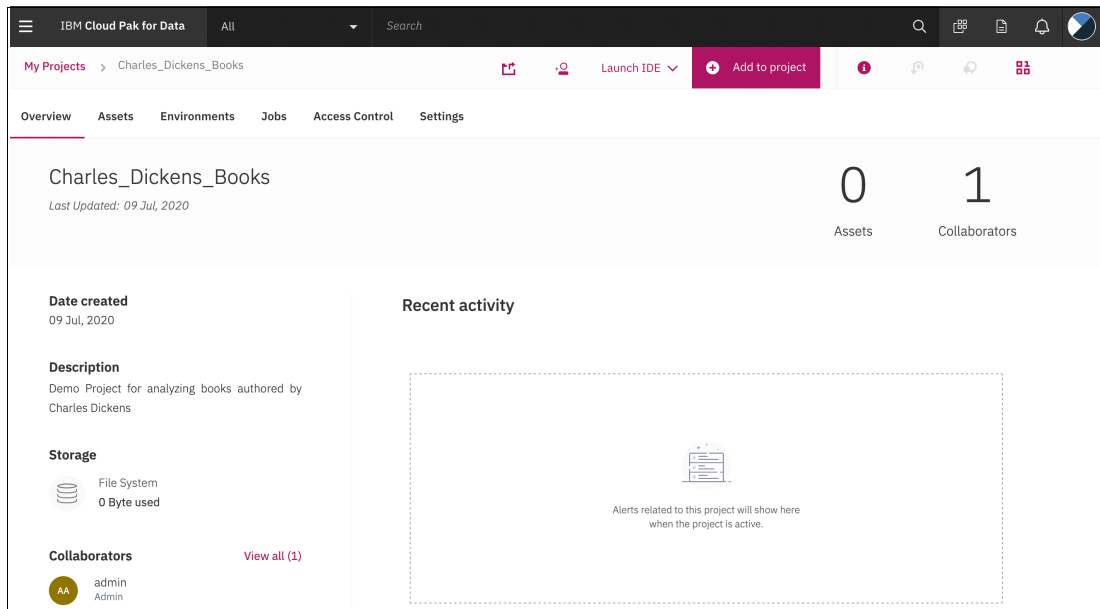


Figure 4-14 Example of the project workspace

6. Go to WKC by clicking the horizontal bars at the upper left and selecting **Organize** → **All Catalogs**.
7. Select the catalog containing the Charles_Dickens data set.
8. Select the **author:Charles_Dickens** tag.

9. Select up to 10 documents, and click **Add to Project**. Figure 4-15 shows an example.

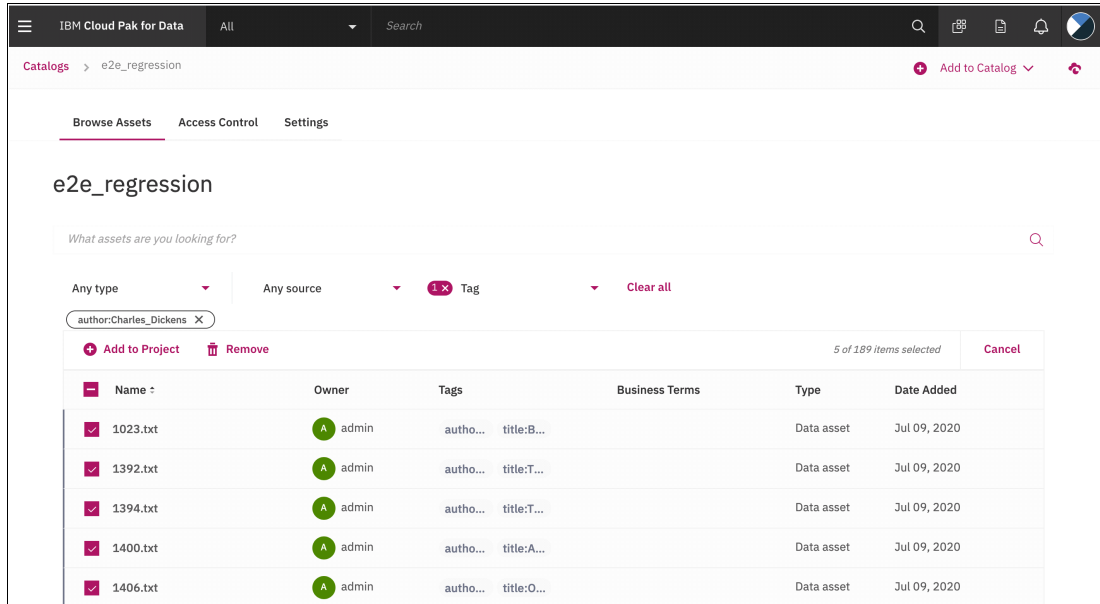


Figure 4-15 Select up to 10 documents and click Add to Project

10. Select the **Charles_Dickens_Books** project and click **Add**.

11. Go to the **Charles_Dickens_Books** project. Click the horizontal bars at the upper left, select **Projects**, and select the **Charles_Dickens_Book** project.

12. Click the **Assets** tab, and verify that the assets are available in the project. Figure 4-16 shows an example.

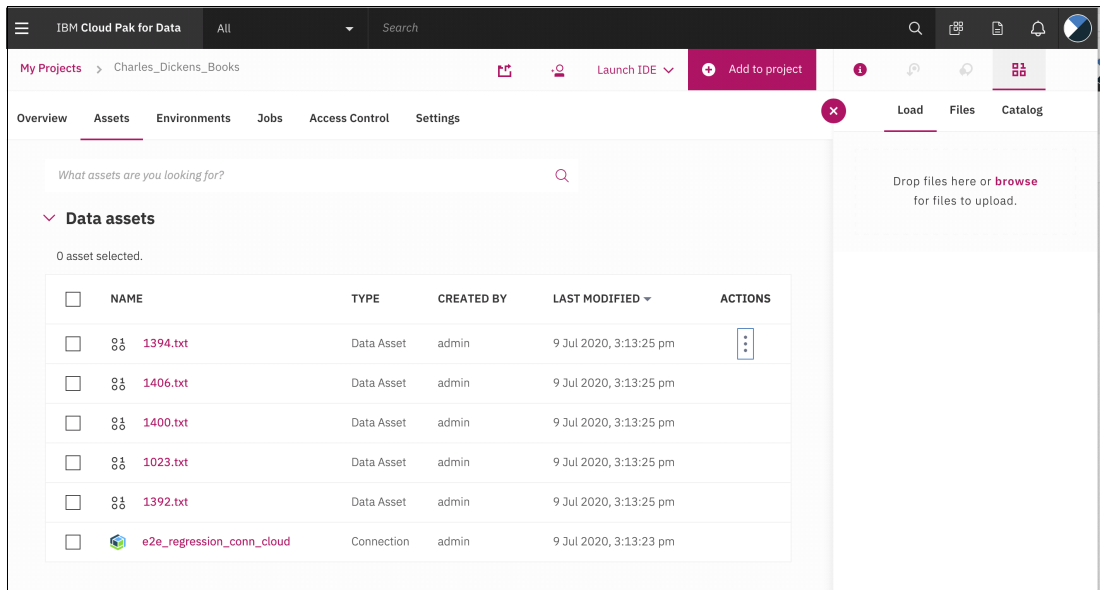


Figure 4-16 Verifying that the assets are available in the project by clicking the Assets tab

Within the project space, you can use the tools that are built into IBM CP4D, such as Spark, Rstudio, and Python. Optionally, click the **Environments** tab or **Launch IDE**.

4.2.3 Creating data governance policies

Data governance policies in WKC control access to the data based on the classification capabilities of IBM Spectrum Discover. By using the data governance Policy Manager tool in WKC, users can create data governance policies to control access to assets based on the tags that are associated with assets that are registered in WKC catalogs by IBM Spectrum Discover. For more information, see 6.3, “Creating a data governance policy in WKC” on page 80.



Healthcare and life sciences use cases

This chapter describes healthcare and life sciences use cases for using IBM Spectrum Discover with IBM Watson Knowledge Catalog (WKC).

This chapter includes the following topics:

- ▶ Generic healthcare use case
- ▶ COVID-19 use case
- ▶ Breast cancer use case

5.1 Generic healthcare use case

In this use case, a major medical center wanted to better organize and manage research and clinical trial data by addressing four key elements:

1. Cataloging a large genomic reference data set.
2. Monitoring and reporting on data location.
3. Finding Protected Health Information/Personally Identifiable Information (PII) data from genomic and medical imaging data sets.
4. Establishing data usage patterns.

IBM Spectrum Discover solved their problem, and the customer is rolling out 30 PB of IBM Spectrum Discover capacity that is being used to analyze and develop more use cases and insight into the 100+ PB of medical data stored online.

IBM Spectrum Discover can provide tremendous value to the healthcare industry because it simplifies artificial intelligence (AI) enabled solutions from Collect to Infuse, and leverages healthcare data to AI.

5.1.1 IBM Spectrum Discover large-scale AI and data governance with Watson Knowledge Catalog

In this example, IBM Spectrum Discover integrates with the WKC component of IBM Cloud Pak for Data (IBM CP4D) to enrich the catalog content in IBM Spectrum Discover along with the associated data that is available in WKC and IBM CP4D. From an end to end IBM solution point of view, IBM CP4D and WKC support streaming data and clinical data from structured databases. IBM Spectrum Discover complements this support by adding support for unstructured data, including built-in content inspection support for healthcare / life sciences file types, such as genomics .vcf and .bam files, and Digital Imaging and Communications in Medicine (DICOM) medical images.

This solution is illustrated in Figure 5-1 on page 51.

IBM Spectrum Discover Integration with Watson Knowledge Catalog

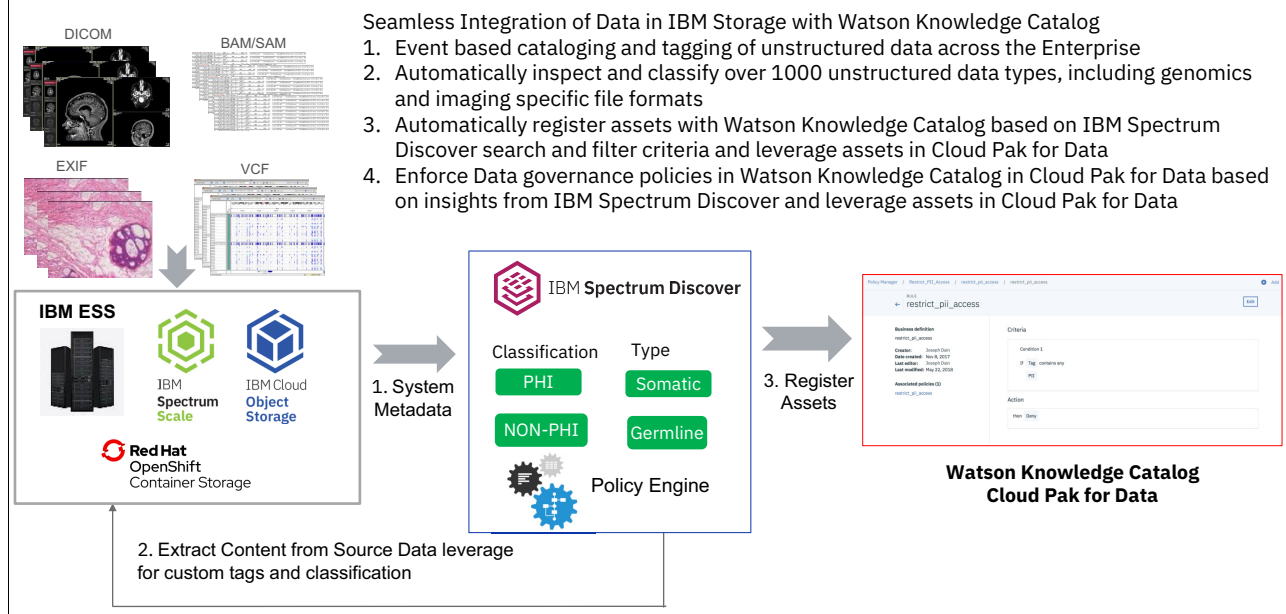


Figure 5-1 IBM Spectrum Discover and Watson Knowledge Catalog for healthcare and life sciences

Using this architecture, the medical center focused on two basic usages for IBM Spectrum Discover:

- ▶ Data governance
- ▶ Large scale-analytics, AI, and machine learning (ML)

5.1.2 Data governance: Medical file classification example

Integration of IBM Spectrum Discover and WKC provides governance tools for regulations and compliance requirements and restrictions on information access and data in silos.

In this example, an organization has unstructured data that might contain sensitive information in the DICOM headers of medical images in one or more data sources. A regular expression (regex) is created to determine the value of the Patient Identity Removed DICOM headers, as shown in Figure 5-2.

Add Regular Expression

Name
DICOM-PatientIdentityRemoved

Description
DICOM Patient Identity Removed Header

Regular Expression Pattern
^.*Patient Identity Removed[s+([A-Z]+):+].*\$

Cancel Save Expression

Figure 5-2 Regular expression to extract a patient identity removed header

Next, an IBM Spectrum Discover policy is created by using the regex to populate the value of PatientIdentityRemoved with the value from the DICOM header, as shown in Figure 5-3.

Add new policy

Inactive Active

Name
DICOM-Patient-Information-Removed

Policy Type
CONTENT SEARCH

Collections
Type search collection

Filter
path like '/ibm/scale0/watson_health%'

Application
contentsearchagent

Tag
PatientIdentityRemoved

Search Expression
Search Expression
DICOM-PatientIdentityRemoved

Value
Select a value

Schedule
 Now Daily Weekly Monthly

Save Cancel

Figure 5-3 DICOM Patient Identity Removed classification policy

Next, medical images that have the PatientIdentityRemoved tag set to YES are selected and registered in WKC in IBM CP4D to conduct a study and collaborate with fellow researchers. Figure 5-4 shows the data set search and export user experience.

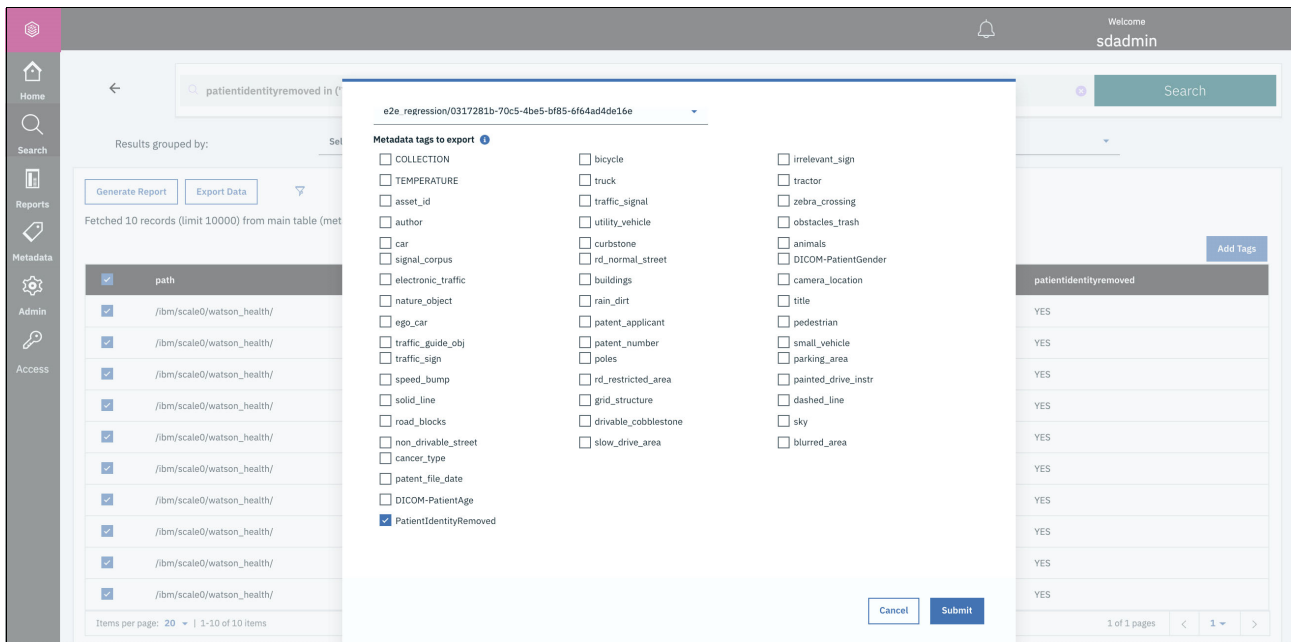


Figure 5-4 Exporting DICOM Images to Watson Knowledge Catalog from IBM Spectrum Discover

WKC active policy enforcements capabilities can also be leveraged to control access to information about the assets.

5.1.3 Large-scale analytics, AI, and ML for healthcare and life sciences

In this example, a medical researcher wants to find all medical images for patients ages 56 and above to conduct a new experiment leveraging IBM CP4D. In this case, an IBM Spectrum Discover content inspection policy is created to extract the patient age from the DICOM headers that are embedded in the DICOM images. The results are automatically indexed into the IBM Spectrum Discover catalog. A search is performed to export the data set to WKC in IBM CP4D to leverage the suite of AI and analytics tools in the IBM CP4D platform.

Figure 5-5 illustrates this workflow.

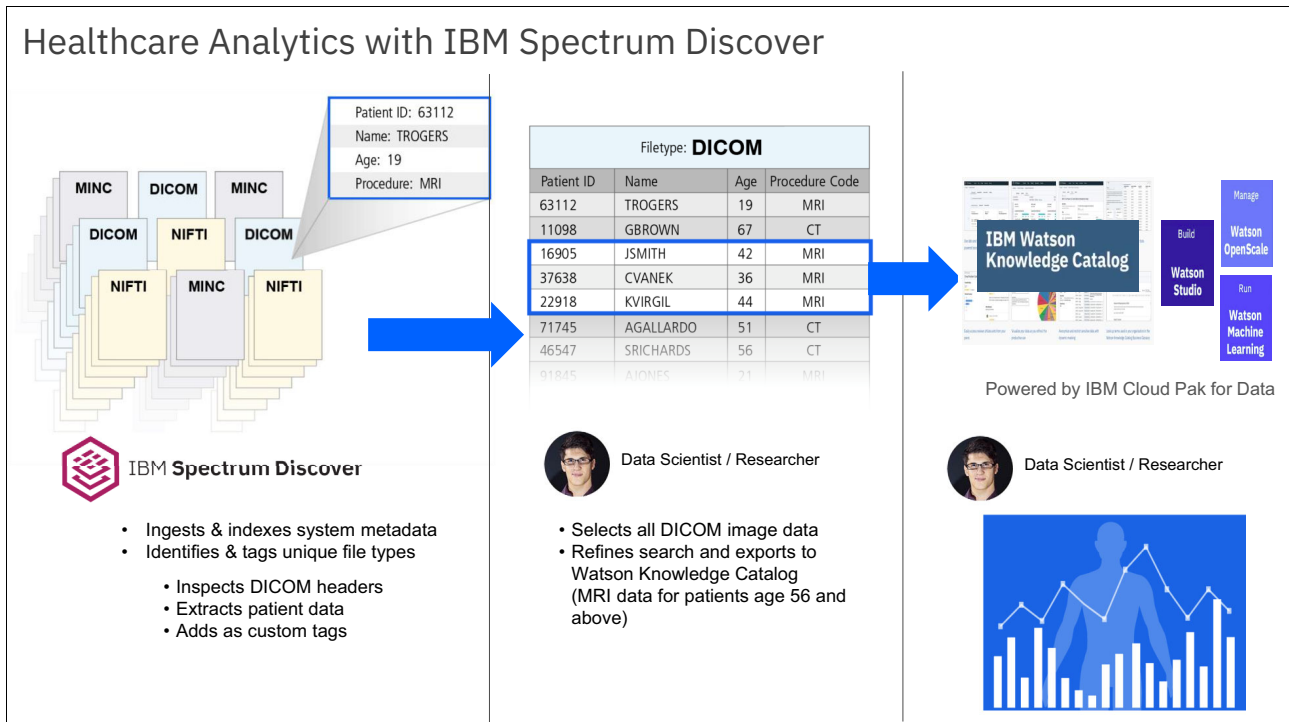


Figure 5-5 Healthcare analytics with IBM Spectrum Discover

Similar to the policy that is described in Figure 5-3 on page 52, an IBM Spectrum Discover content extraction policy that leverages a regex to extract the Patient's Age DICOM header field is run to populate the IBM Spectrum Discover database with this information. Figure 5-6 shows the regex that is used in this example.

Edit Regular Expression

Name
DICOM-PatientAge

Description
DICOM Patient Age

Regular Expression Pattern
^*Patient[']s Age\s+[A-Z]+\s+(\.s.*)\$

Figure 5-6 Patient's Age regular expression

5.2 COVID-19 use case

The data set that is used for this demonstration (Figure 5-7) is a collection of x-ray images of the lungs of various conditions, such as bacterial pneumonia, viral pneumonia, and COVID-19.

Note: What is COVID-19?





- ▶ Coronavirus (COVID-19) is an illness that is caused by a virus that can spread from person to person.
- ▶ The virus that causes COVID-19 is a new or *novel* coronavirus that has spread throughout the world.
- ▶ COVID-19 symptoms can range from mild (or no symptoms) to severe illness.^a

a. <https://www.cdc.gov/coronavirus/2019-ncov/faq.html>

COVID-19 Dataset - Building the IBM Spectrum Discover Catalog

Enable searching for signals in the data

- Identify low score inference results and use to retrain model for higher accuracy
- Population Studies
- Search by patient ID

COVID-19 	Pneumonia Bacteria 
Normal 	Pneumonia Virus 

IBM Visual Insights Inference Model Metadata

Category <ul style="list-style-type: none">• COVID-19• Pneumonia bacteria• Pneumonia virus• Normal	Score <ul style="list-style-type: none">• Accuracy of inference
--	--

Contextual Metadata

Patient ID <ul style="list-style-type: none">• Maps samples to patients	Additional Metadata <ul style="list-style-type: none">• Gender• Age• Location (hospital)• Modality (x-ray)• Survival (yes / no)• Intubation present (yes / no)
--	--

30 June 2020/ © 2018 IBM Corporation


Partial contextual metadata: <https://github.com/ieee8023/covid-chestxray-dataset> 

Figure 5-7 IBM Spectrum Discover catalog for COVID-19 lung x-ray use case

A computer vision inference model leveraging IBM Visual Insights was created to classify these images and detect COVID-19 in them. Other types of computer vision inference models leveraging TensorFlow and BM Watson Machine Learning Accelerator can be used in this AI pipeline.

The COVID-19 data set contains some contextual metadata, such as the age of the patients, the gender of the patients, the location where they were treated, whether they were intubated, and whether they survived. This additional contextual information is leveraged along with the inference model labels to provide extra input and signals about the data to derive new types of results.

One example (Figure 5-8) is that we can quickly look at all the COVID-19 samples of the x-ray images and map that data against age, gender, and location to understand what types of trends that we can derive from the data.

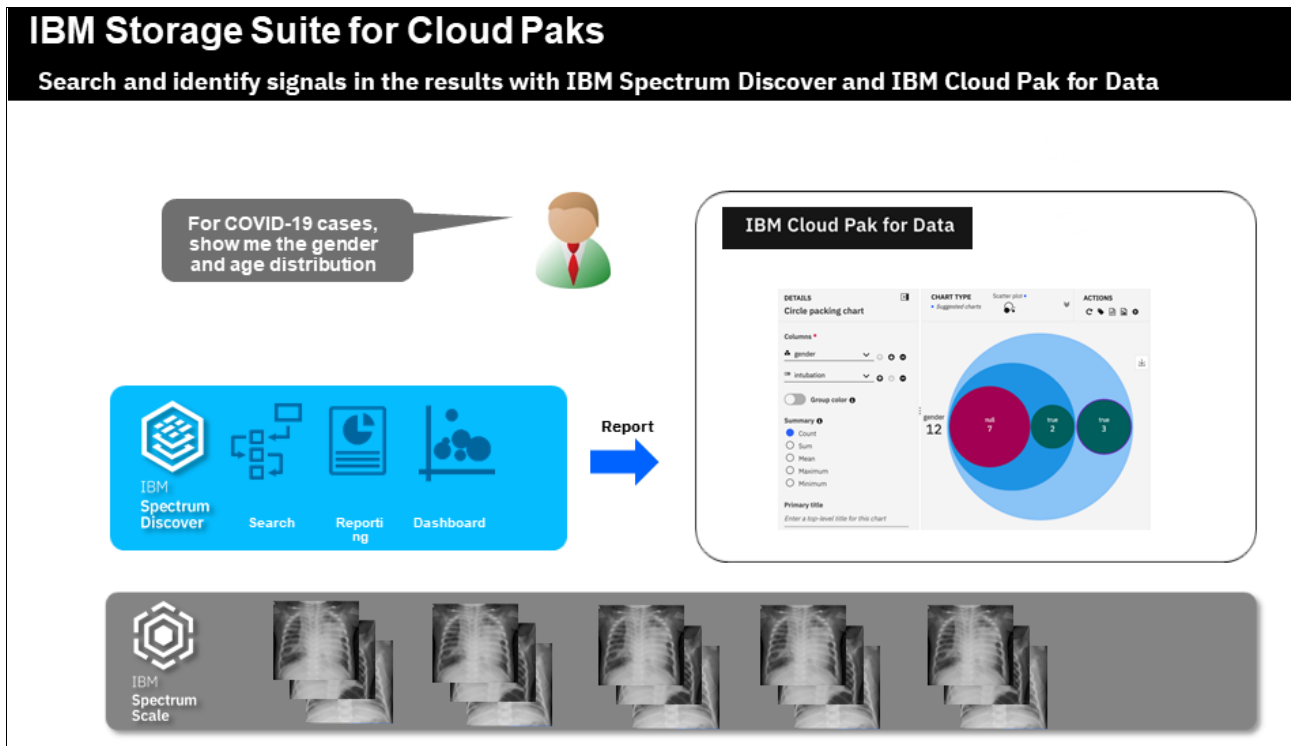


Figure 5-8 IBM Spectrum Discover and IBM CP4D COVID-19 gender and age distribution

5.2.1 Classifying images with IBM Visual Insights

From a detailed workflow point of view, the x-ray images are ingested into IBM Spectrum Scale, and then are cataloged into IBM Spectrum Discover, as shown in Figure 5-9 on page 57.

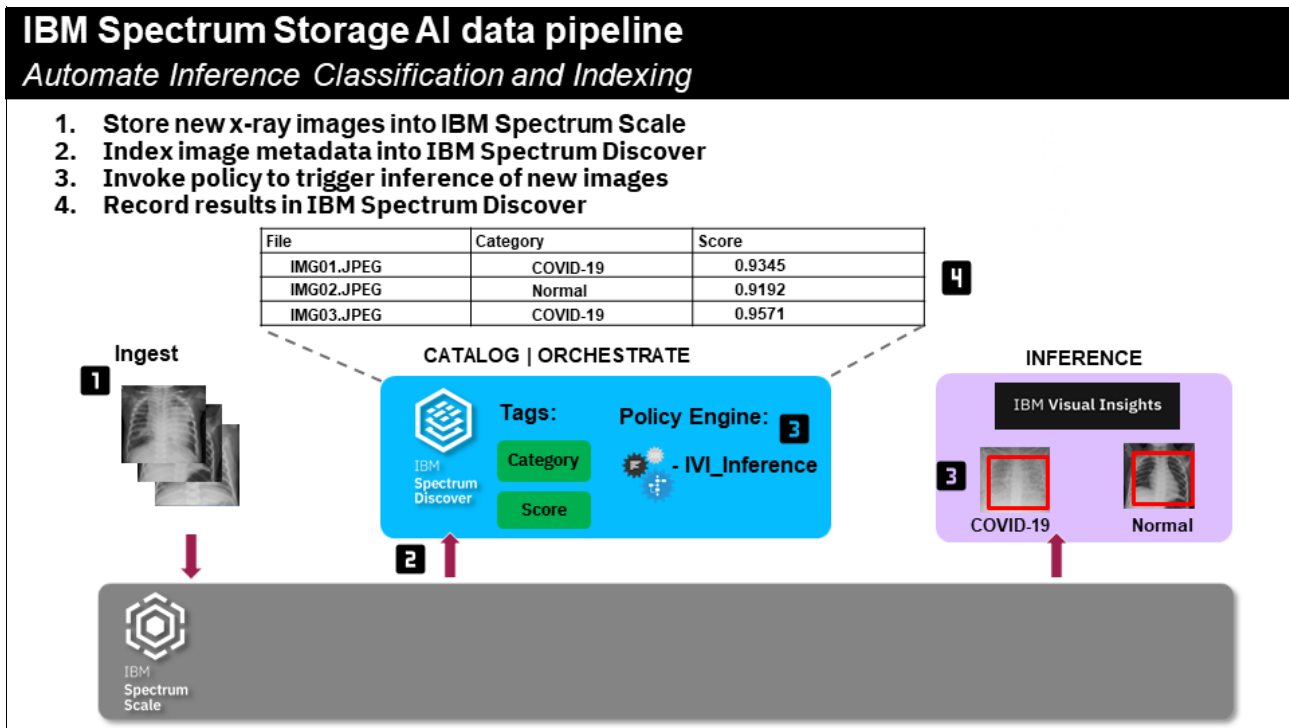


Figure 5-9 IBM Spectrum Storage AI data pipeline: Automate Inference Classification and Indexing

IBM Spectrum Discover automatically invokes a policy to read these images and send them to the computer vision inference model and capture the output, which is then put into the IBM Spectrum Discover catalog and made searchable, as shown in Figure 5-10.

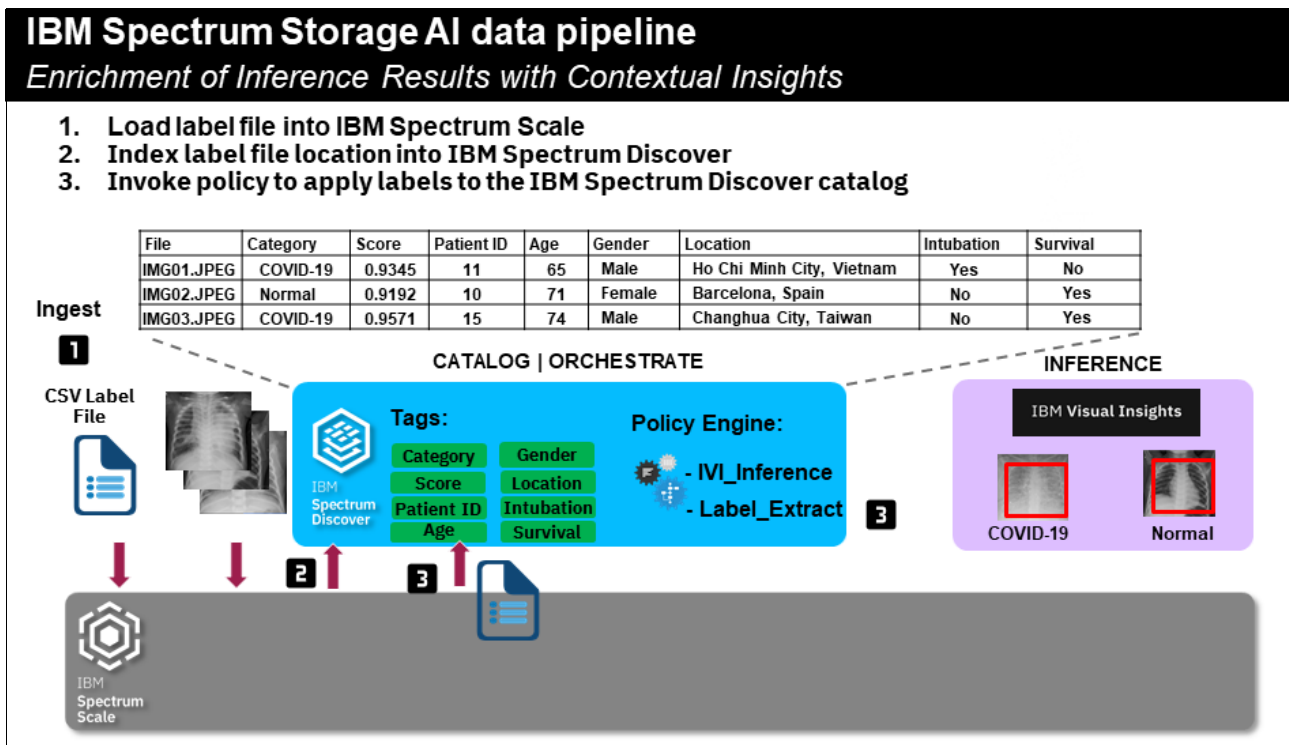


Figure 5-10 IBM Spectrum Storage AI Data Pipeline: Enrichment of Inference Results with Contextual Insights

Then, this inference information is enriched with more contextual information, such as a patient ID patient name and gender, so that we get a 360-degree view of the x-ray samples from this particular data set. The enriched results of the x-ray image inferencing along with the contextual metadata is automatically registered into WKC, where the ecosystem of analytics and AI tools is used to do further analysis on the data.

5.2.2 Registering assets and tags / labels into Watson Knowledge Catalog

In IBM Spectrum Discover, various data sets are cataloged across the heterogeneous storage environment. We can view the catalog of data sets and select the COVID-19 data set within this data set to get a breakdown of the distribution in the different disease categories, such as COVID-19, bacterial pneumonia, and viral pneumonia, and the results of disease-free x-rays, as shown in Figure 5-11.

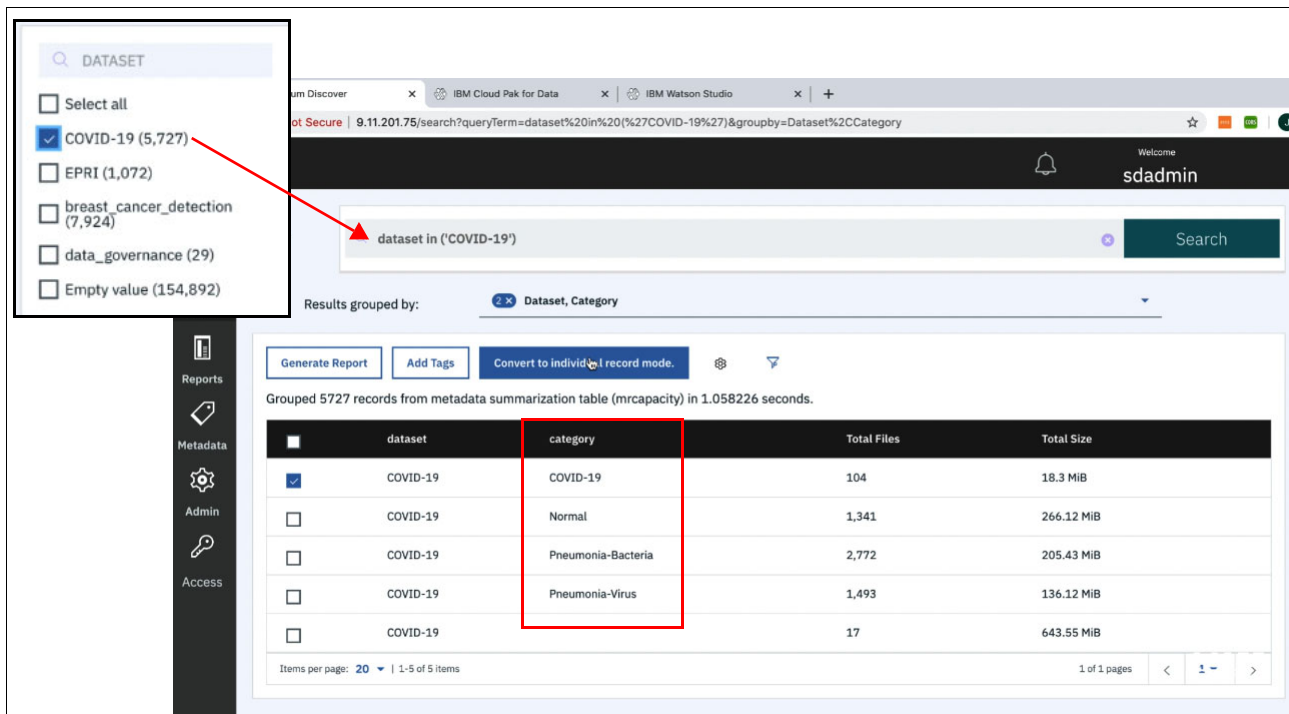


Figure 5-11 IBM Spectrum Discover COVID-19 data set with disease category

We can then select the x-ray images containing COVID-19 and export them directly into WKC. In this example, we select a handful of these images. When exporting to WKC, we include custom tags that were derived from the data set. In this case, we select the data set name and a disease category tag, as shown in Figure 5-12 on page 59.

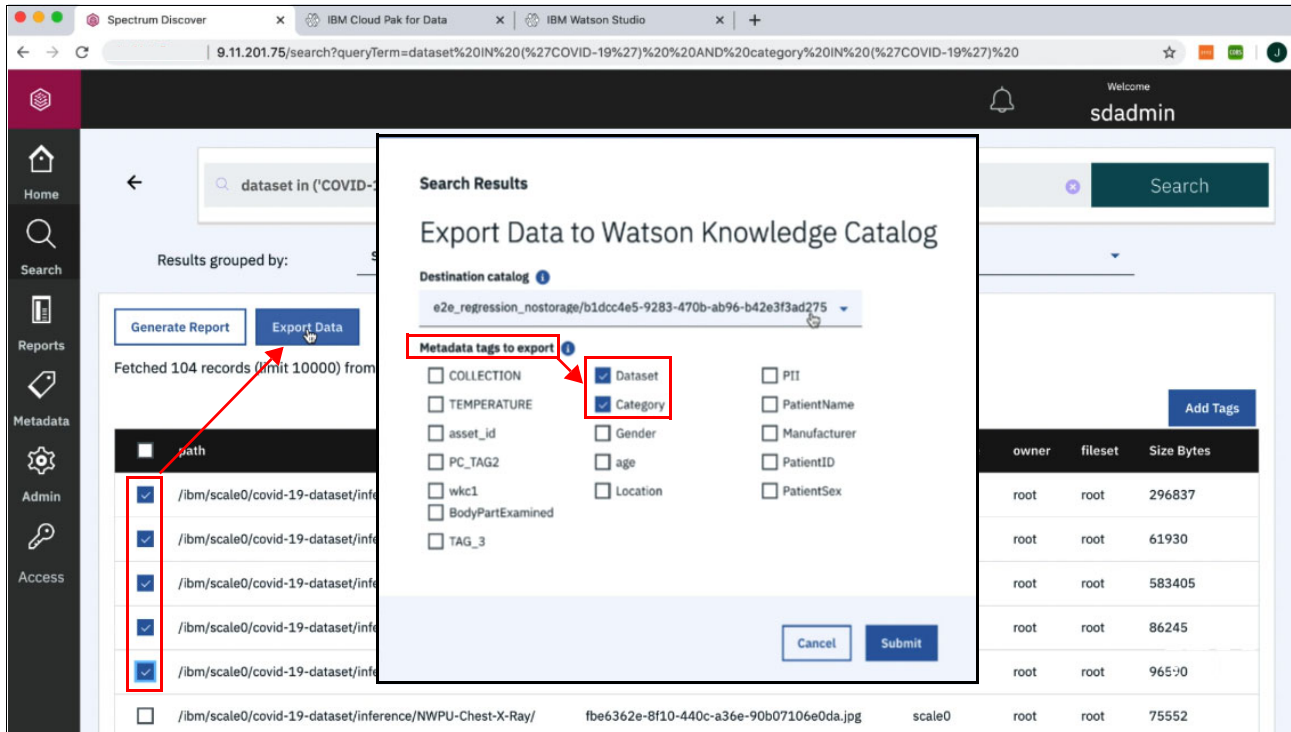


Figure 5-12 Selecting and exporting COVID-19 records to Watson Knowledge Catalog

5.2.3 Viewing images in Watson Knowledge Catalog

Next, we select some x-ray images of normal lungs and export them to the WKC. Again, we select the COVID-19 data set. This time, we select images that have the Normal category. A handful of these images to match the normal disease category is selected and then exported to WKC. The data set name along with the disease category is also persisted to WKC.

Figure 5-13 shows this process.

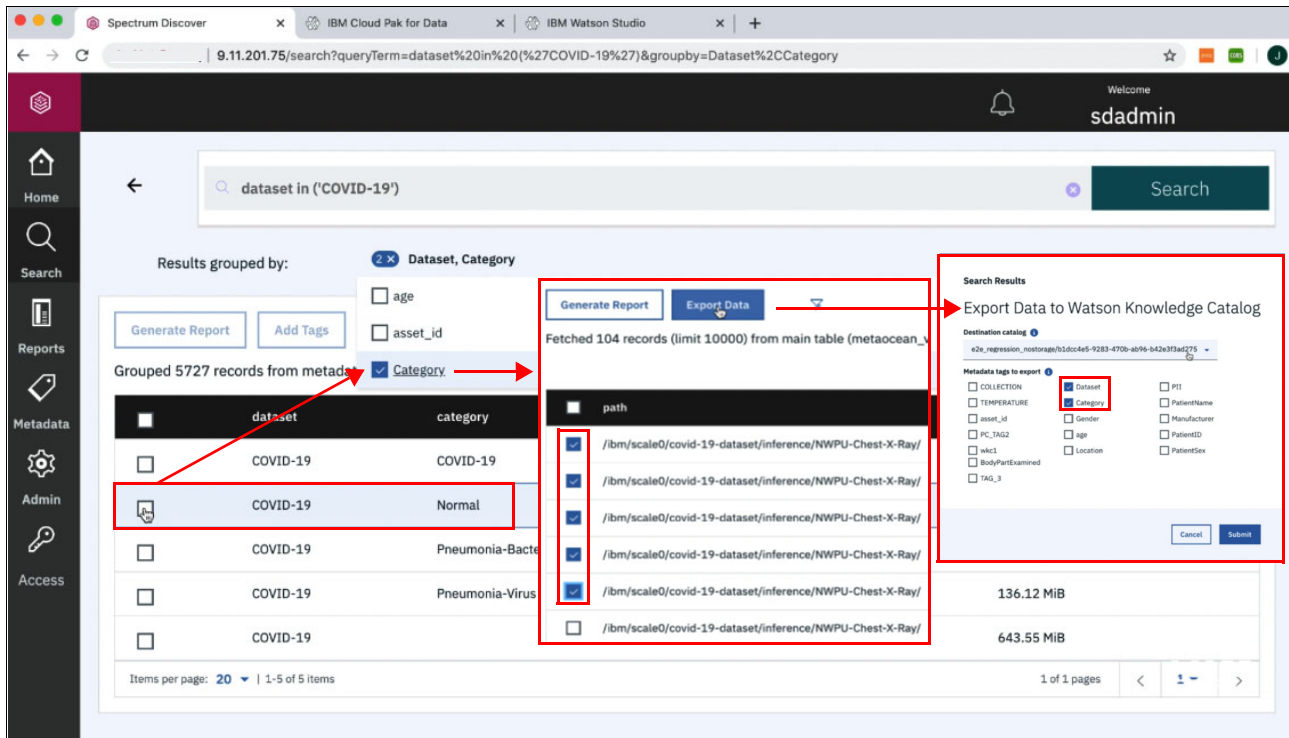


Figure 5-13 Selecting Normal for the disease category and the records to export to Watson Knowledge Catalog

Now that the content is integrated within WKC in IBM CP4D, we see that we have various different x-ray images that are registered by IBM Spectrum Discover. We also see the tags that are associated with these x-ray images, such as the disease category and the data set name. We select images that are tagged as normal. Then, we preview these images within the WKC.

Figure 5-14 shows the selection of the Normal Category data.

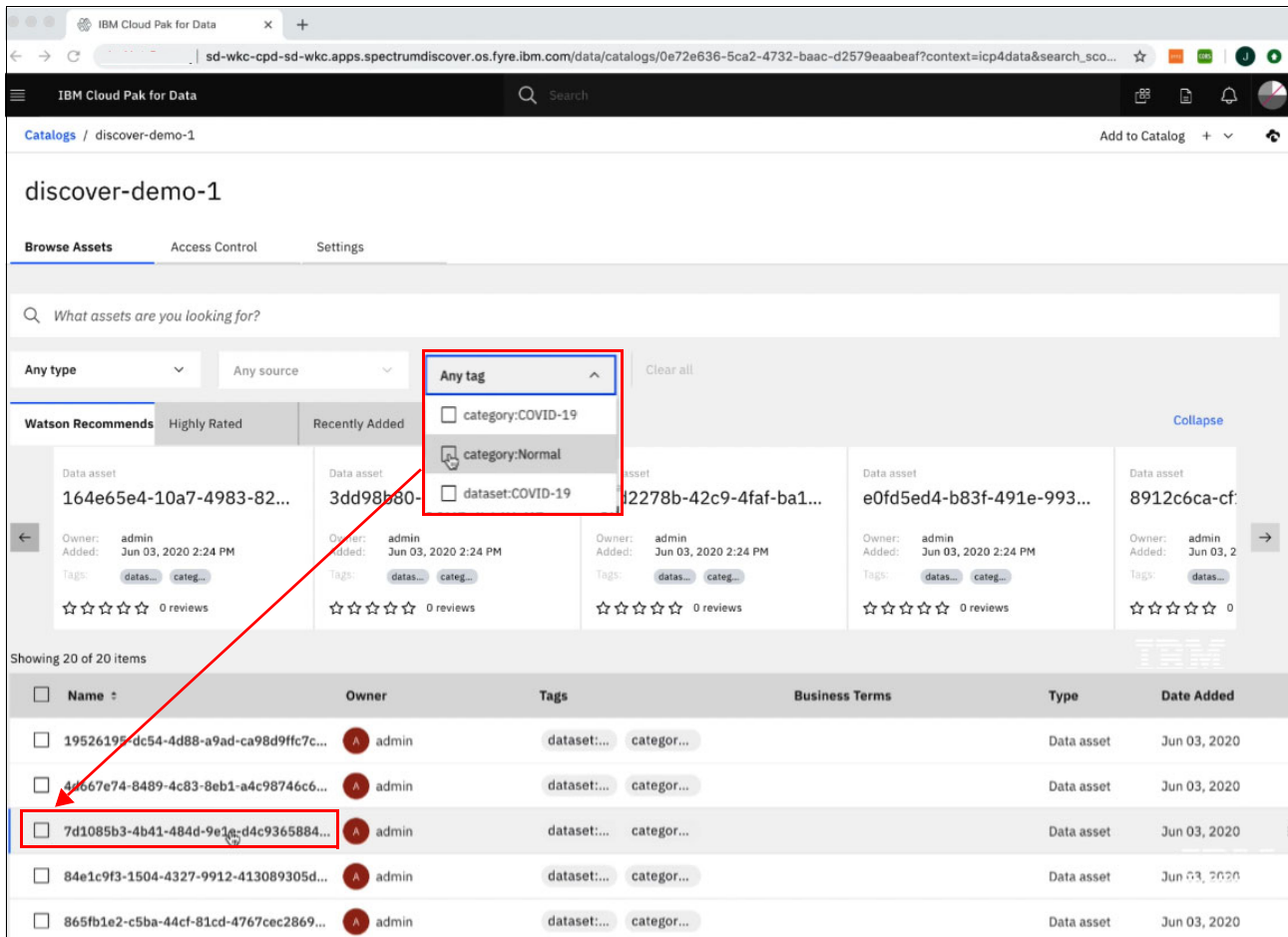


Figure 5-14 Selecting the Normal Category tag in IBM CP4D

Figure 5-15 shows the image of a normal lung x-ray.



Figure 5-15 Displaying a normal lung image in IBM CP4D

By selecting the COVID-19 category tag, we see in Figure 5-16 an x-ray of a COVID-19 lung.

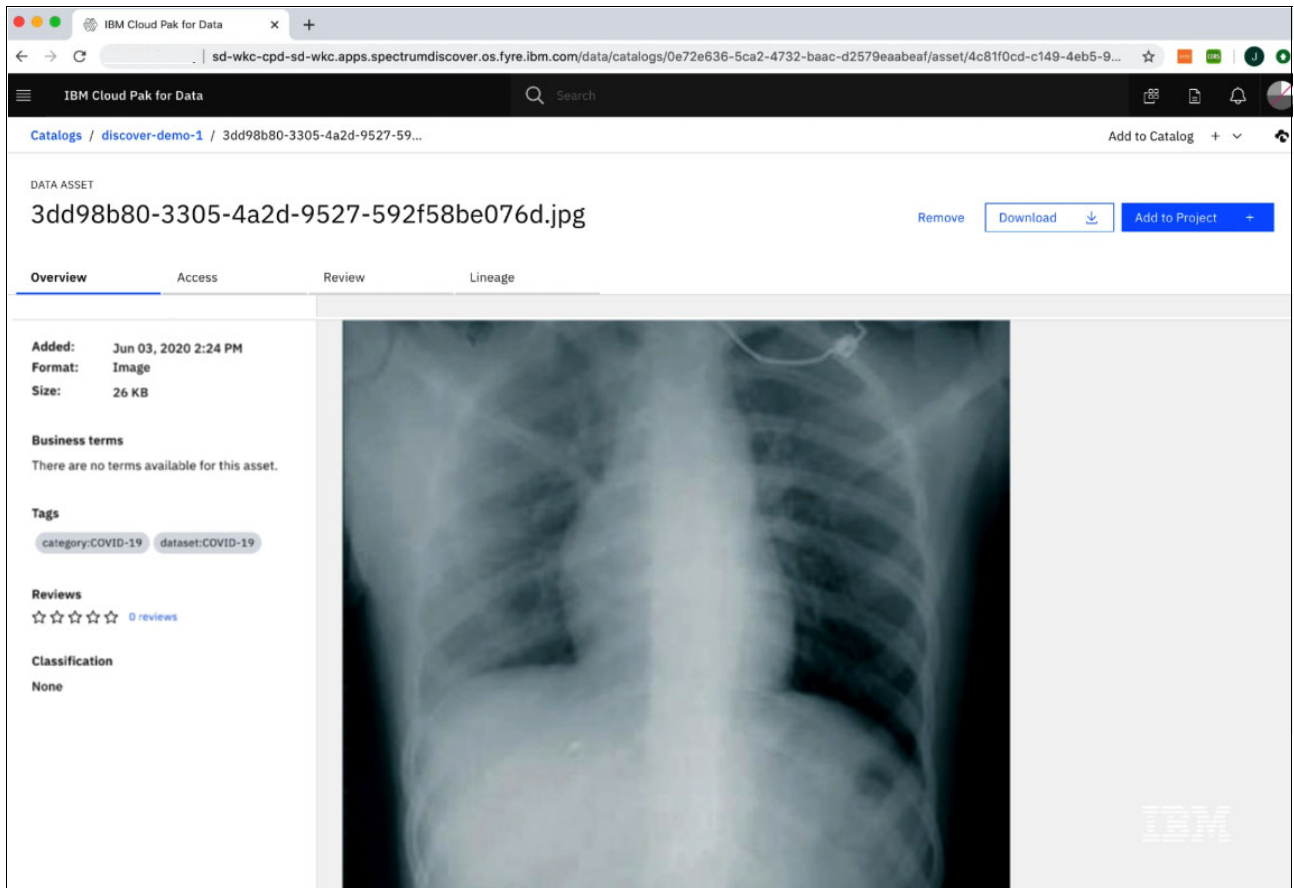


Figure 5-16 Lung x-rays containing COVID-19

These images can be leveraged by extra tools in IBM CP4D to do further analysis. For example, we can instantiate an IBM Watson ML instance (Figure 5-17) and use these images to train a new model.

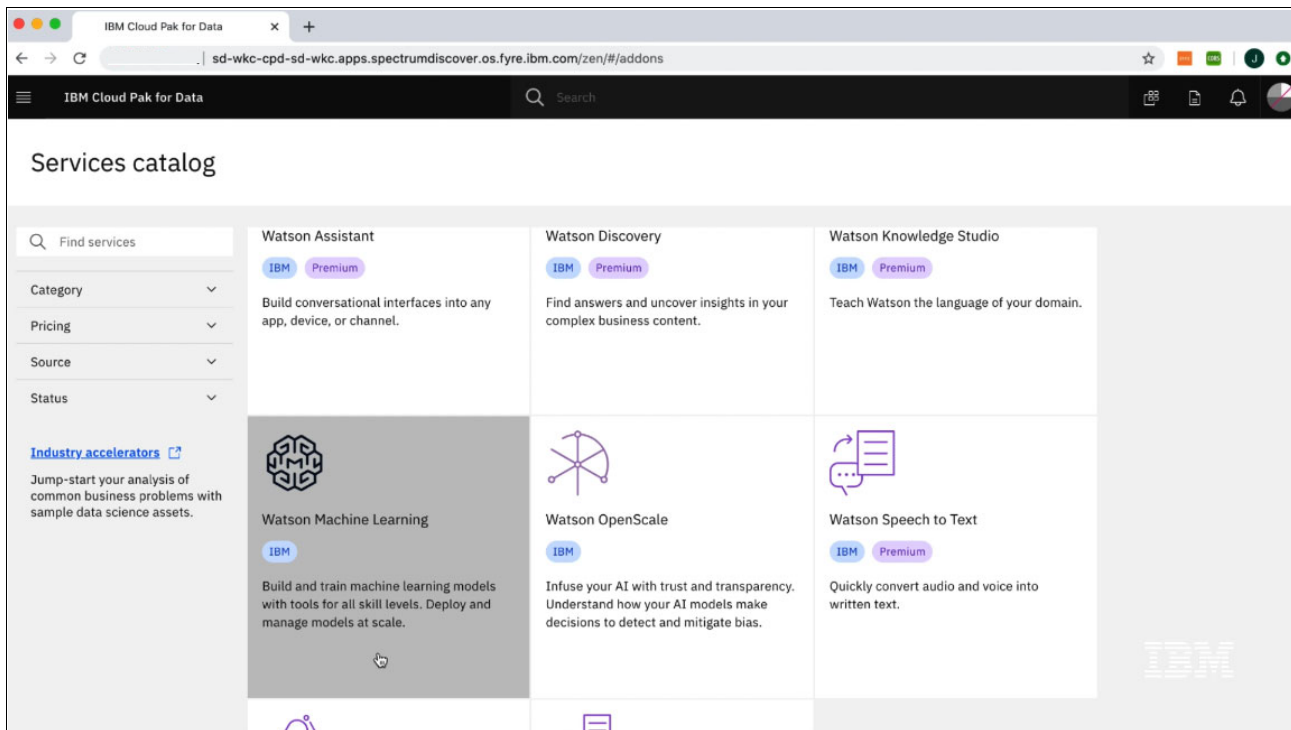


Figure 5-17 Watson Machine Learning in IBM CP4D can be used to train new COVID-19 models

5.2.4 Uploading an IBM Spectrum Discover custom report into Watson Knowledge Catalog

You can leverage the additional contextual information that is provided by IBM Spectrum Discover to get more insight about the data. For example, the COVID-19 data set contains more information, such as the age of the patient, the gender of the patient, and the location where the patient was treated. By combining this extra information, you can start to see signals in the data.

For example, Figure 5-18 on page 65 shows that there are patients ages 50 and above that were treated in Spain, both male and female, that have COVID-19.

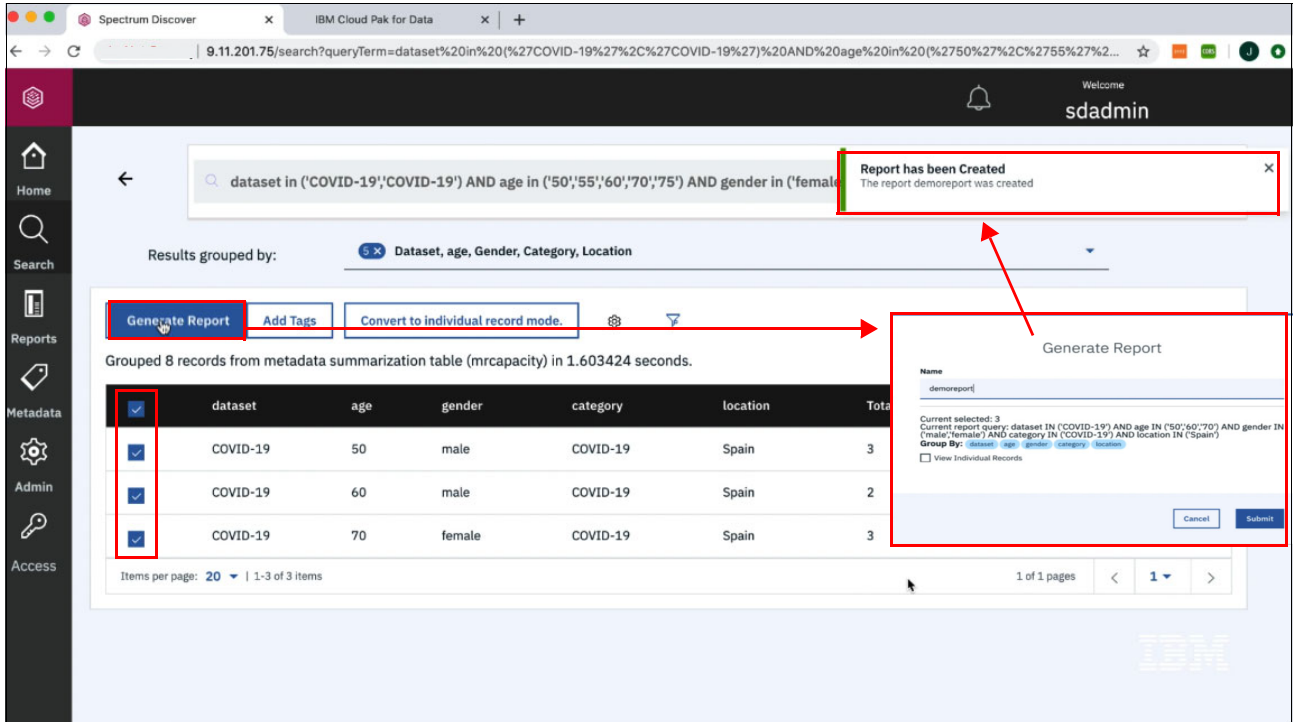


Figure 5-18 Creating a report with more information from the COVID-19 data set

You can select this data set and generate a manifest file and a report, which can be imported into WKC to do more visual analytics. In this case, we preinstalled the IBM Spectrum Discover manifest file and a report into WKC and IBM CP4D, and use some of the visualization tools to better understand the results.

More specifically for the COVID-19 patients that are shown in Figure 5-19, we look at their gender and whether these patients were intubated. For this image, we see that for males that there are two samples where they were intubated, and for females we see that for all three samples they were intubated.

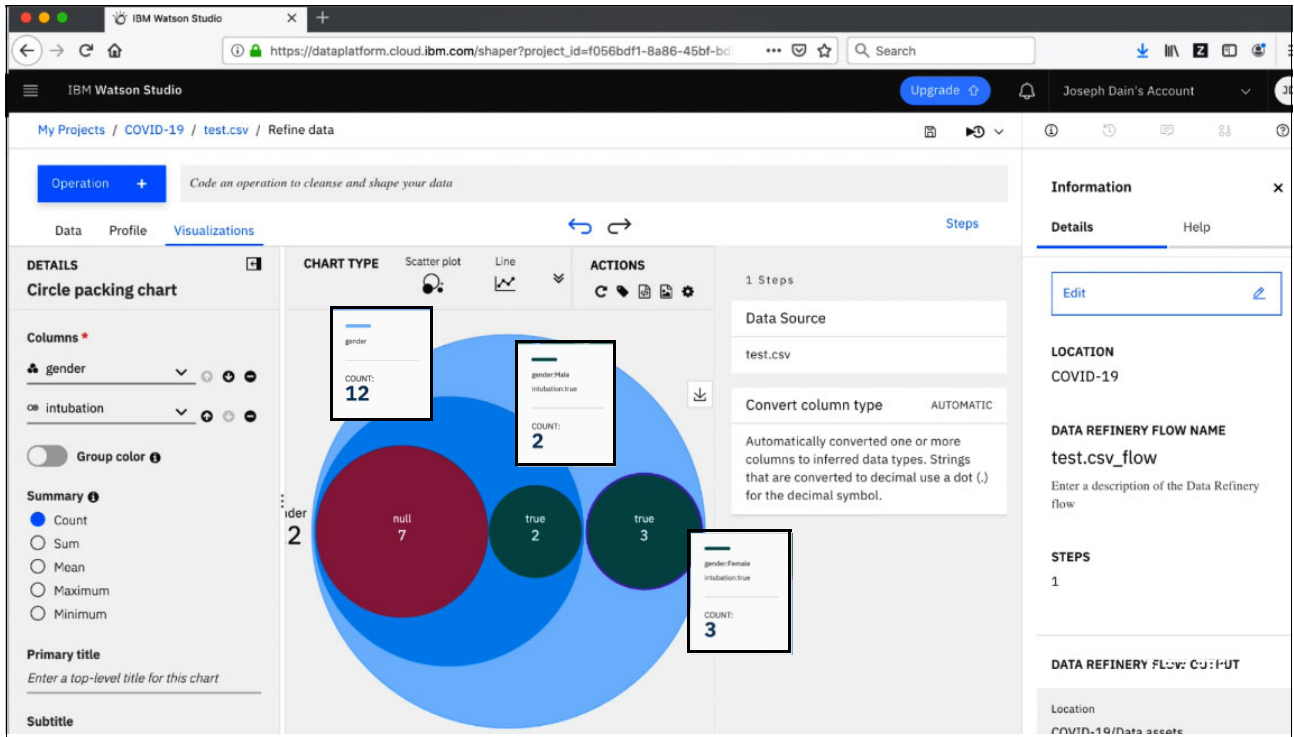


Figure 5-19 IBM Watson Studio Word Cloud chart showing gender and intubation

We can also look at the age distribution of the COVID-19 samples that are mapped against gender, as shown in Figure 5-20 on page 67. In this case, we can see that most of the samples are from males with the highest age being 75, the lowest being 50, and the median age being 55.

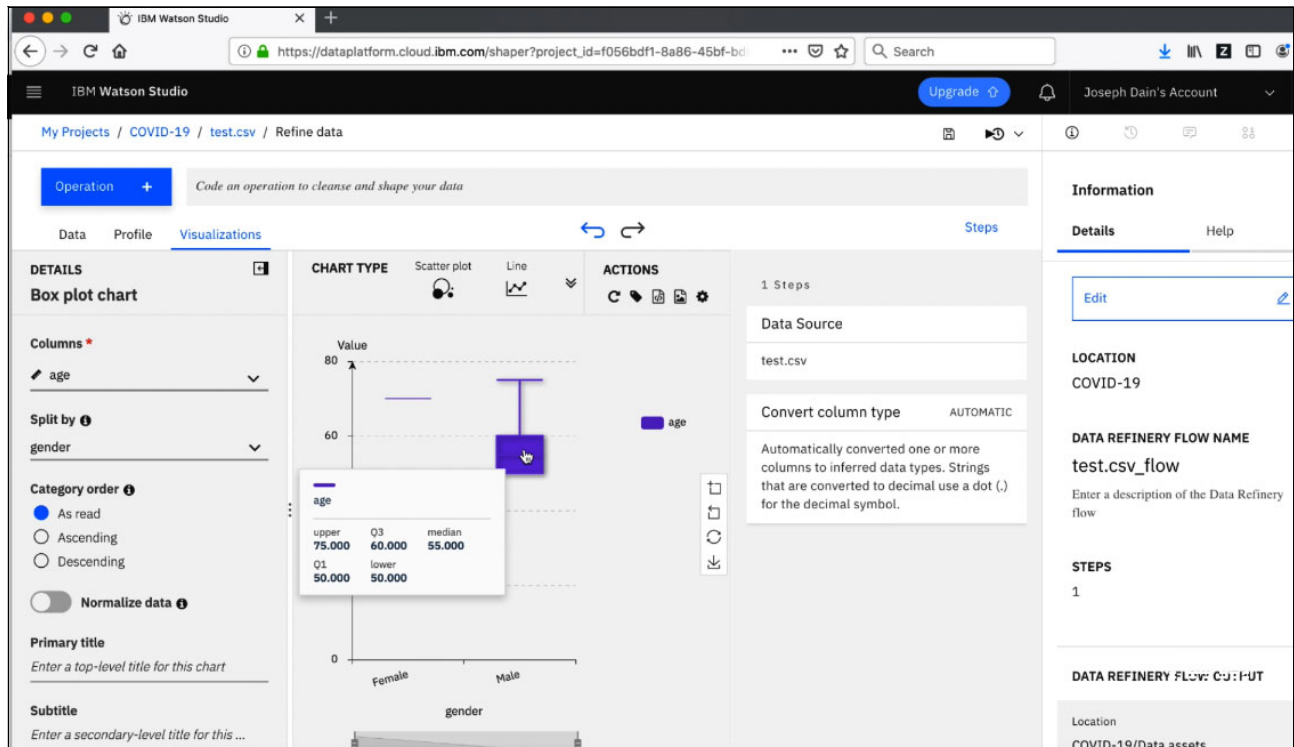


Figure 5-20 IBM Watson Studio box plot chart of COVID-19 age and gender data

By leveraging these types of tools, you can gain more insight into the data and images. IBM Spectrum Discover also provides insight into vast amounts of unstructured data that can be registered with WKC.

5.3 Breast cancer use case

This section describes a use case for breast cancer research that shows how you can explore data that was harvested by IBM Spectrum Discover and displayed graphically in a Jupyter Notebook. A public data set containing breast cancer images, which is made available by the Federal University of Paraná in Brazil, was used to illustrate one of the use cases in this book. A complete description of the data and how to request a download can be found at [Breast Cancer Histopathological Database \(BreakHist\)](#).

The Breast Cancer Histopathological Image Classification is composed of 9,109 microscopic images of breast tumor tissue that was collected from 82 patients by using different magnifying factors (40X, 100X, 200X, and 400X). For the use case in the book, we selected the 1,820 images with a magnification of 400X. The use case contains 2,480 benign and 5,429 malignant samples (700X460 pixels, 3-channel RGB, 8-bit depth in each channel, and PNG format).

Both benign and malignant breast tumors can be sorted into different types based on the way the tumoral cells look under the microscope. Various types and subtypes of breast tumors can have different prognoses and treatment implications. The data set contains four histological distinct types of benign breast tumors: adenosis (A), fibroadenoma (F), phyllodes tumor (PT), and tubular adenoma (TA), and four malignant tumors (breast cancer): carcinoma (DC), lobular carcinoma (LC), mucinous carcinoma (MC), and papillary carcinoma (PC).

Images typically contain headers, which contain metadata that can be useful to the business. There is considerable work in image annotation. Image annotation can be done with IBM Spectrum Discover. This process is described in detail in section 3.3.2 “Digital Imaging and Communications in Medicine use case” of *IBM Spectrum Discover: Metadata Management for Deep Insight of Unstructured Storage*, REDP-5550, which describes the details of extracting the medical metadata from DICOM formatted images. These images were annotated by the BreakHist by using the Backus-Naur Form (BNF). For more information about the BNF, see [IBM Knowledge Center](#).

Each image file name stores information about the image itself: method of procedure biopsy, tumor class, tumor type, patient identification, and magnification factor.

For example, SOB_B_TA-14-4659-40-001.png is image 1, at magnification factor 40X, of a benign tumor of type TA, originally from slide 14-4659, which was collected by a Surgical Open Biopsy (SOB) procedure.

The extra medical metadata in this case is derived from the full path of each file. For example:

```
/<root>/malignant/SOB/ductal_carcinoma/SOB_M_DC_14-10926/400X/SOB_M_DC-14-10926-400-001.png
```

There is also metadata that is extracted from the POSIX file attributes, such as size, creation date, and owner:

```
-rw-rw-r-- 1 ibm ibm 468651 Feb 27 2015 SOB_M_DC-14-10926-400-001.png
```

This information is automatically extracted by IBM Spectrum Discover. Image metadata is shown in Table 5-1, and more metadata is shown in Table 5-2.

Table 5-1 IBM Spectrum Discover image metadata descriptions

Image metadata	Explanation	Values
Tumor Type	Type of tumor	Benign / Malignant
Cancer Type	Type of cancer	fibroadenoma, mucinous_carcinoma, tubular_adenoma, adenosis, papillary_carcinoma, lobular_carcinoma, ductal_carcinoma, phyllodes_tumor
Patient ID	Patient ID number	<number>
Magnification	Resolution of image	40X, 100X, 200X, 400X

Table 5-2 IBM Spectrum Discover extra metadata descriptions

Extra metadata	Explanation	Values
Path	Directory path in file system	</path to file/File>
Filename	Name of file	File name
Owner	File system owner	Owner ID
Access	Access control list	Read - Write

Extra metadata	Explanation	Values
Timestamps	Time file was created, last modified, and last accessed in the file system.	[Datetime]
Size	File size	Size in Bytes

The path has the class malignant, type ductal carcinoma, magnification, and patient ID. The information is also coded into the name of the file.

Using IBM Spectrum Discover tagging, we mapped the information in the file path to information tags, as shown in Figure 5-21.

Tag Name	Type	Value
TEMPERATURE	Open	
email	Characteristics	
vcf_format	Open	
vcf_project_format	Open	
vcf_project_reference	Open	
SOB	Open	
cancer_type	Open	
tumor_type	Open	
patient_id	Open	
magnification	Open	
custom_tag	Open	
dr_level	Open	

Figure 5-21 IBM Spectrum Discover Tags

Figure 5-22 lists the tags that are created when the policies are run.

Policy ▲	Type	Schedule (UTC)	Status	Progress	Collections	Last Modified by
vcf_format	CONTENTSEARCH	Done	active stopped	100% 0 failed out of 46		sdadmin
Extract_tumor_type	AUTOTAG	Done	active stopped	100% 0 failed out of 7909		sdadmin
Extract_dr_level	AUTOTAG	Done	active stopped	100% 0 failed out of 35126		sdadmin
set_temperature	AUTOTAG	Done	active stopped	100% 0 failed out of 653430	spectrum-discover	sdadmin
Tumor_Type	AUTOTAG	Done	active stopped	100% 0 failed out of 14		sdadmin
emailpolicy	CONTENTSEARCH	Done	active stopped	100% 0 failed out of 517401		sdadmin
extract_vcf4_reference	CONTENTSEARCH	Done	active stopped	100% 0 failed out of 2		sdadmin
extract_vcf32_program	CONTENTSEARCH	Done	active stopped	100% 0 failed out of 44		sdadmin
SOB	AUTOTAG	Done	inactive stopped	0%		sdadmin
Extract_cancer_type	AUTOTAG	Done	active stopped	100% 0 failed out of 5429		sdadmin
Extract_Patient_Id	AUTOTAG	Done	active stopped	100% 0 failed out of 7909		sdadmin

Figure 5-22 Tags that are created when the policies are run

Figure 5-23 on page 71 shows that the tag "cancer_type" is mapped to the seventh field in the path name. The "%" represents a wildcard, and matches anything in the field.

```
[ 'path', 'filename', 'filetype', 'datasource', 'owner', 'group', 'revision',
  'site', 'platform', 'cluster', 'inode', 'permissions', 'fileset', 'uid', 'gid',
  'recordversion', 'state', 'migloc', 'mtime', 'atime', 'ctime', 'tier', 'size',
  'fkey', 'collection', 'temperature', 'duplicate', 'sizeconsumed', 'nodename',
  'filespace', 'mgmtclass', 'vcf_format', 'vcf_project_format',
  'vcf_project_reference', 'sob', 'cancer_type', 'tumor_type', 'patient_id',
  'magnification', 'custom_tag', 'dr_level', 'email' ]
```

Edit policy

Inactive Active

Name **Policy Type**

collections
Type search collection

Filter
path like '/a9000/BreakHis_v1/histology_slides/breast/malignant/SOB/%/%/'

Extract tag from path

Tag Name **Depth**

Schedule
 Now Daily Weekly Monthly

Figure 5-23 Tag "cancer_type" is mapped to the seventh field in the path name

5.3.1 Using Data Refinery, Jupyter Notebook, or Cognos to analyze report data

IBM Spectrum Discover is a software layer that enables access to storage back ends through application programming interfaces (APIs). Users can point their favorite front-end client or interface to IBM Spectrum Discover and access the metadata that is related to the files or objects.

In this example, Jupyter Notebook, a popular user interface for data science, was used, with a Python Version 3 kernel, to access this metadata directly in IBM Spectrum Discover. The following IBM Spectrum APIs were used:

- ▶ [Authentication API example](#)
- ▶ [Report reader API example](#)

The data was read into a Python Panda Dataframe, which is a data structure that is convenient for data manipulation.

Figure 5-24 a picture of the data structure, highlighting some of the chosen metadata to display cancer type, tumor type, and patient ID.

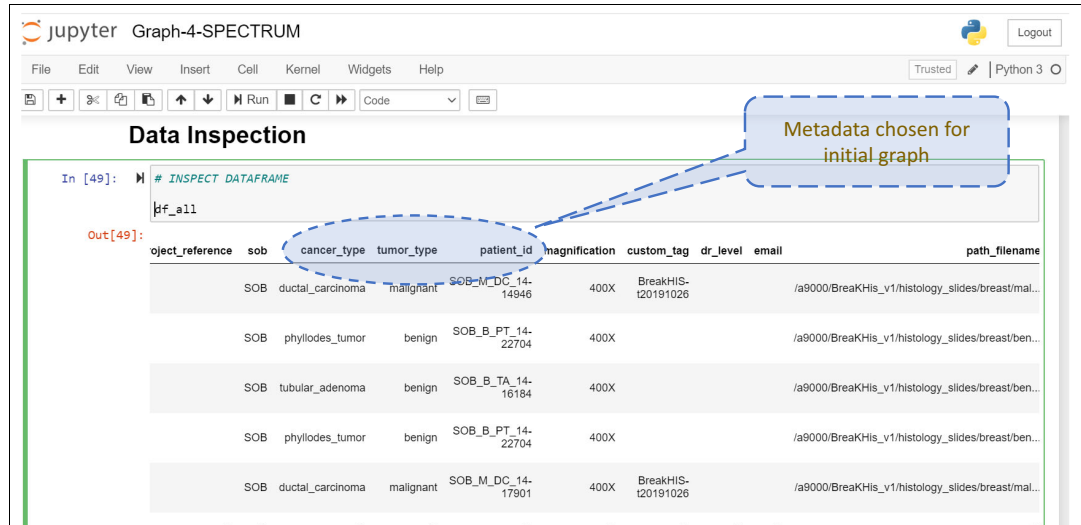


Figure 5-24 Data Inspection in Jupyter Notebook

After data of interest is identified, it can be displayed in many ways. A convenient and powerful technology for analysis and visualization of genomic data is graph theory. Basically, a graph consists of *nodes* that represent some type of entity, and *edges*, which represent links or relationships between those entities. Nodes and edges can have attributes or properties. In this example, patient ID and cancer type were chosen as nodes, with an edge labeled 'diagnosed' linking them:

Patient ID -> diagnosed -> cancer type

In addition, tumor type was set as an attribute of cancer type.

The Python library `networkx` was used to create the graph directly from the data frame.

To display the graph, we took advantage of a convenient and interactive open source display library that was recently published by the University of California San Diego School of Medicine. It is named [visJS2Jupyter](#).

visJS2Jupyter enables an interactive view of graph data. Figure 5-25 on page 73 shows the types of display that are obtained after basic manipulation of the library parameters. The cancer types are represented by circles, with their size proportional to the number of patients, and can be conveniently color-coded.

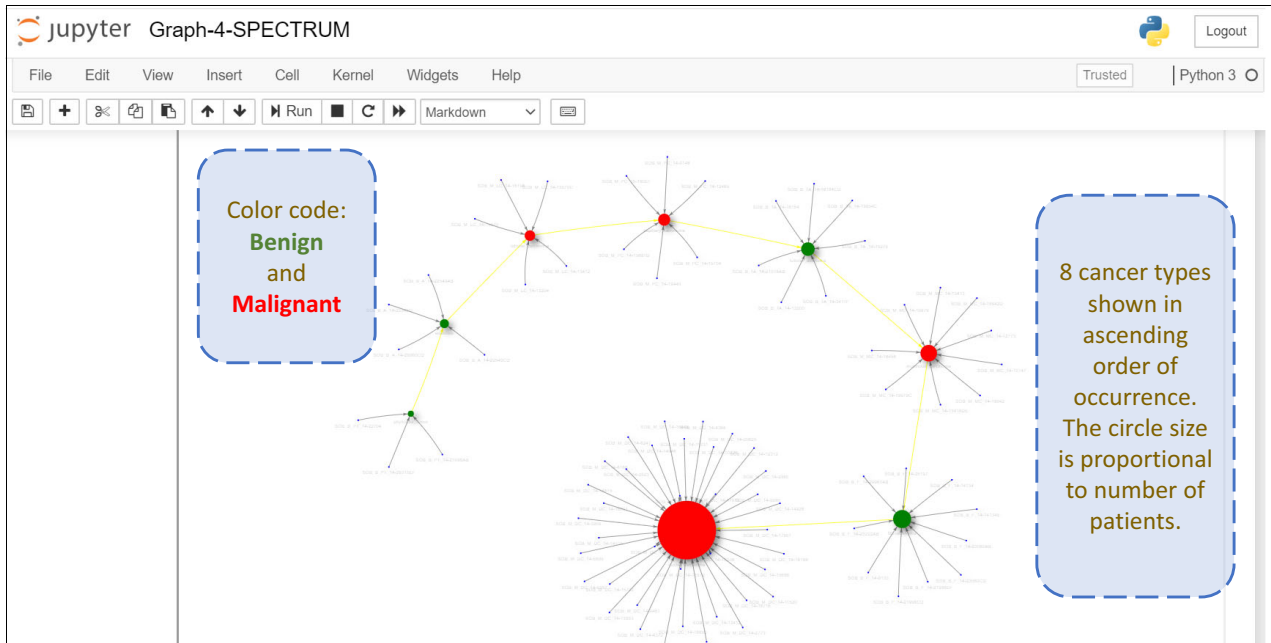


Figure 5-25 Graph Display in Jupyter Notebook: High level

The graph can be manipulated by the user, who can, for example, zoom in and out, or click and drag elements for more clarity. By zooming in, as shown in Figure 5-26, the user can make the patient ID nodes readable.

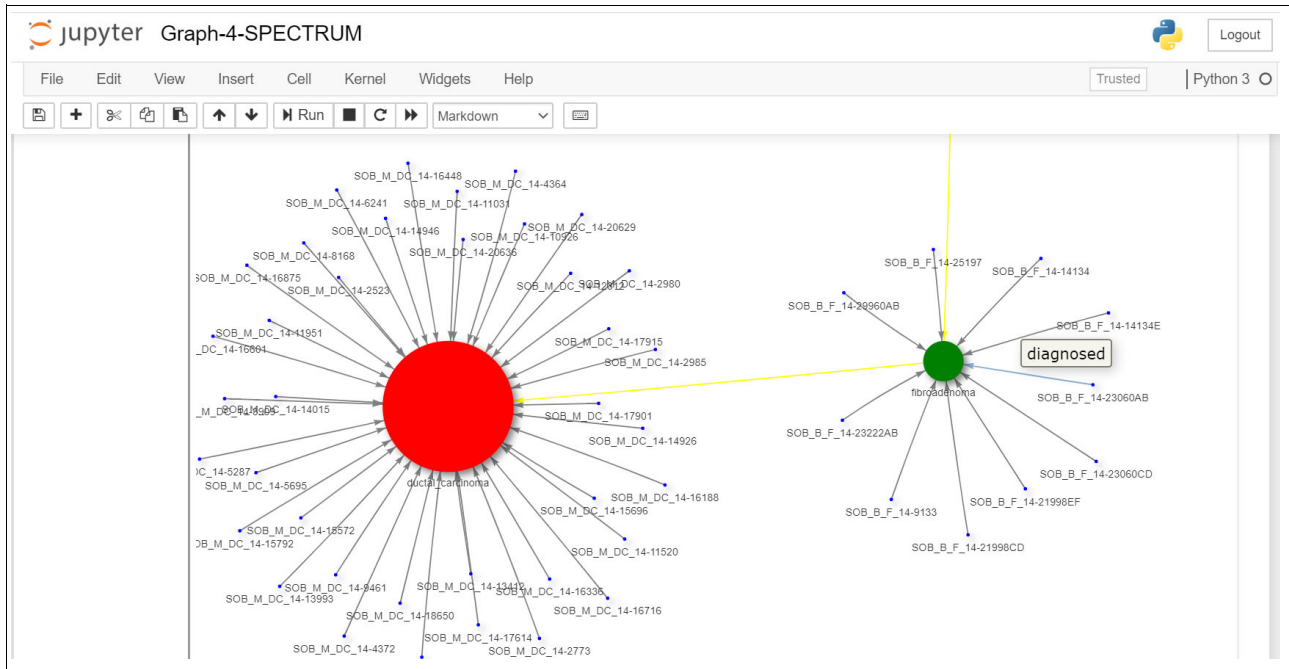



Figure 5-26 Graph display in Jupyter Notebook: Detail

The interactivity of the graph display enables the user to browse the display and discover information such as attributes. In Figure 5-26, the user clicks an edge and sees the attribute 'diagnosed' linking the particular patient ID (SOB_B_F_14-23060AB) and the cancer type (fibroadenoma).



Financial services use case: Personally Identifiable Information detection and data governance

This chapter describes a financial services use case for Personally Identifiable Information (PII) detection and data governance by using IBM Spectrum Discover with IBM Watson Knowledge Catalog (WKC).

This chapter includes the following topics:

- ▶ Current challenges in financial industries
- ▶ Protecting cardholder data with PCC DDS use case
- ▶ Creating a data governance policy in WKC

6.1 Current challenges in financial industries

The financial and banking industries are facing unprecedented change as they move toward digital transformation and digitalization. Although most investors, bankers, and financial customers embrace the technological revolution, there are still challenges that must be overcome.

The future of financial service and banking requires new ideas and methods for accomplishing tasks on a greater scale. The customer will be at the forefront of the future. Today's financial and banking customer expects more, demands faster access, and expects better results than in the last 5 years. Banks and financial institutions that cannot compete with these expectations will likely struggle to maintain viability in the end.

6.1.1 Customer expectations

The customer experience is at the forefront of the challenges facing the banking industry today. In many ways, traditional banks are not delivering the level of service that customers are demanding, especially when it comes to technology. For example, more customers are using mobile devices for transactions. A 2020 study found that more than 60 percent of financial, investor, and banking customers use their smartphones or other mobile devices. But, customers still expect in-person customer service too. The same study found that 25 percent would not be comfortable opening an account with a bank that did not have a local presence.

6.1.2 Increasing pressure from competition

Young customers especially are open to change in their financial services provider. In a recent survey, Accenture found that 31 percent of banking customers would consider banking with Facebook, Amazon, or Google if they offered the same type of services they currently enjoy. Already, financial technology startups like Robinhood or Acorns are taking advantage of this mindset by offering apps that support investing and other innovative financial services.

6.1.3 Investor expectations

Despite all the news about banking profits, banks and other financial institutions are not meeting their shareholders' expectations for return on investment or equity. Part of the reason is the lack of accurately understanding customer expectations, which translate into lower customer enrollment and retention rates.

6.1.4 Keeping up with compliance and regulations

Regulations in the financial service and banking industry continue to increase. Banks and financial institutes are spending a large part of their income on making sure that they are compliant with and follow government regulations. They must make sure that there are systems in place to keep up with ever-changing regulations and industry standards.

Traditional banks constantly evaluate and improve their operations to keep up with fast-changing consumer and shareholder expectations, technology, and industry regulations.

The following list describes some of the key attributes that play major roles for financial service firms addressing regulatory challenges:

- ▶ Geopolitical change: Companies must expect business change and disruption.
- ▶ Divergent regulation: You must anticipate continued differences in state, federal, and global regulations among protectionist and localized public policy agendas in the US and overseas.
- ▶ Data protection and governance: Protect your data at all costs.
- ▶ Operational resilience: Plan for the unexpected. It happens.
- ▶ Credit quality: Firms must apply what they have learned from past credit cycles.
- ▶ Capital and liquidity shifts: Even though there might be an easing of regulatory capital and liquidity requirements, firms should not weaken risk management.
- ▶ Compliance agility: You must have a solution for agile and streamlined compliance.
- ▶ Financial crime: It is okay to be innovative, but do not increase the risk of financial crime.
- ▶ Customer trust: Firms must maintain the trust of the customers.
- ▶ Ethical conduct: Do the right thing no matter what.

Financial service companies must create a strategy to innovate and stay compliant.

6.1.5 Business agility with the latest technology

Business growth is important for banking and financial firms, but to grow they must spend money updating their technology. According to a report, financial service firms must continue to invest in technology such as robotic process automation and other workflow automation tools to increase their efficiency and reduce the costs that are associated with operations, risk management, and compliance.

Firms must modernize and transform their technology platforms and data storage so they can enable cloud and big data solutions such as artificial intelligence (AI)-supported digital customer support assistants.

Financial firms must also consider consolidating platforms and provide a more efficient, customer friendly experience across internet, mobile, and physical locations.

6.2 Protecting cardholder data with PCC DDS use case

This use case section provides an overview of Payment Card Industry (PCI), PCI requirements, implementing Payment Card Industry Data Security Standard (PCI DDS) into business, and an example of creating a data governance policy in WKC.

6.2.1 Overview of PCI

In today's digital world, everyone participates in payment card transactions, so it is imperative that you use standard security procedures and technologies to thwart theft of cardholder data.

Merchant-based vulnerabilities can appear almost anywhere in the card-processing industry including the following areas:

- ▶ Point-of-sale devices
- ▶ Mobile devices, personal computers, or servers
- ▶ Wireless hotspots
- ▶ Web shopping applications
- ▶ Paper-based storage systems
- ▶ The transmission of cardholder data to service providers
- ▶ Remote access connections

Vulnerabilities can also extend to systems that are operated by service providers and acquirers, which are the financial institutions that initiate and maintain the relationships with merchants that accept payment cards.

Here are some of the PCI Security Standards:

- ▶ **PCI Data Security Standard (PCI DSS)**

The PCI DSS applies to all entities that store, process, or transmit cardholder data. It covers technical and operational system components that are included in or connected to cardholder data. If you accept or process payment cards, PCI DSS applies to you.
- ▶ **Personal Identification Number (PIN) Transaction Security (PTS) Requirements**

The PCI PTS is a set of security requirements that is focused on characteristics and management of devices that are used in the protection of cardholder PINs and other payment processing related activities. The PTS standards include PIN Security Requirements, Point of Interaction (POI) Modular Security Requirements, and Hardware Security Module (HSM) Security Requirements. The device requirements are for manufacturers to follow in the design, manufacture, and transport of a device to the entity that implements it.
- ▶ **Payment Application Data Security Standard (PA-DSS)**

The PA-DSS is for software vendors and others who develop payment applications that store, process, or transmit cardholder data or sensitive authentication data as part of authorization or settlement, when these applications are sold, distributed or licensed to third parties. Most card brands encourage merchants to use payment applications that are tested and approved by the PCI Security Standards Council (PCI SSC).
- ▶ **PCI Point-to-Point Encryption Standard**

This Point-to-Point Encryption (P2PE) standard provides a comprehensive set of security requirements for P2PE solution providers to validate their P2PE solutions, and might help reduce the PCI DSS scope of merchants that use such solutions. P2PE is a cross-functional program that results in validated solutions incorporating the PTS Standards, PA-DSS, PCI DSS, and the PCI PIN Security Standard.
- ▶ **PCI Card Production Logical Security Requirements and Physical Security Requirements**

The Card Production Logical and Physical Security Requirements address card production activities, including card manufacturing, chip embedding, data preparation, pre-personalization, card personalization, chip personalization, fulfillment, packaging, storage, mailing, shipping, PIN printing and mailing (personalized, credit, or debit), PIN printing (non-personalized prepaid cards), and electronic PIN distribution.
- ▶ **PCI Token Service Provider Security Requirements**

The Token Service Provider (TSP) Security Requirements are intended for TSPs that generate and issue Europay, Mastercard, and Visa (EMV) Payment Tokens, as defined under the EMV Payment Tokenization Specification Technical Framework.

6.2.2 Overview of PCI requirements

PCI Security Standards (Figure 6-1) are technical and operational requirements that are set by the PCI SSC to protect cardholder data. The standards apply to all entities that store, process, or transmit cardholder data, with requirements for software developers and manufacturers of applications and devices that are used in those transactions. PCI SSC is responsible for managing the security standards, and compliance with the PCI set of standards is enforced by the founding members of the different councils.

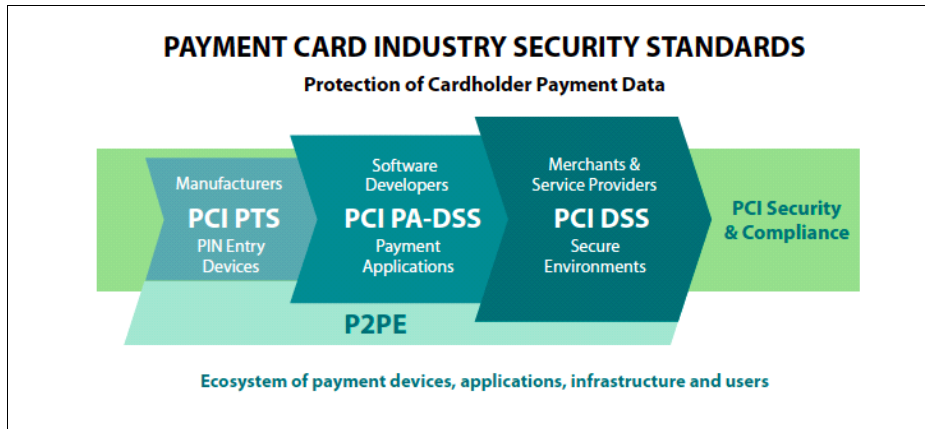


Figure 6-1 Payment Card Industry Security Standards

The PCI Data Security Standard

PCI DSS is the global data security standard that is adopted by the payment card brands for all entities that process, store, or transmit cardholder data or sensitive authentication data. It consists of steps that mirror security best practices.

Table 6-1 PCI DSS goals and requirements

Goals	PCI DSS requirements
Build and maintain a secure network and systems.	<ol style="list-style-type: none"> 1. Install and maintain a firewall configuration to protect cardholder data. 2. Do not use vendor-supplied defaults for system passwords and other security parameters.
Protect cardholder data.	<ol style="list-style-type: none"> 1. Protect stored cardholder data. 2. Encrypt transmission of cardholder data across open, public networks.
Maintain a vulnerability management program.	<ol style="list-style-type: none"> 1. Protect all systems against malware and regularly update anti-virus software or programs. 2. Develop and maintain secure systems and applications.
Implement strong access control measures.	<ol style="list-style-type: none"> 1. Restrict access to cardholder data by business need to know. 2. Identify and authenticate access to system components. 3. Restrict physical access to cardholder data.

Goals	PCI DSS requirements
Regularly monitor and test networks.	<ol style="list-style-type: none"> 1. Track and monitor all access to network resources and cardholder data. 2. Regularly test security systems and processes.
Maintain an information security policy.	Maintain a policy that addresses information security for all.

6.2.3 Implementing PCI DSS into business

To ensure that security controls continue to be properly implemented, PCI DSS should be implemented into business-as-usual activities as part of an entity's overall security strategy. This configuration enables an entity to monitor the effectiveness of its security controls on an ongoing basis and maintain its PCI DSS-compliant environment in between PCI DSS assessments.

Here are items that you must review when implementing PCI DSS:

- ▶ Monitoring of security controls to ensure that they are operating effectively and as intended.
- ▶ Ensuring that all failures in security controls are detected and responded to in a timely manner.
- ▶ Reviewing changes to the environment before completion of the change to ensure that the PCI DSS scope is updated and controls are applied.
- ▶ Performing a formal review of the impact to PCI DSS scope and requirements because of changes to your organizational structure.
- ▶ Performing periodic reviews and communications to confirm that PCI DSS requirements continue to be in place and personnel are following secure processes.
- ▶ Reviewing hardware and software technologies at least annually to confirm that they continue to be supported by the vendor and can meet the entity's security requirements including PCI DSS, and remediating shortcomings.

6.3 Creating a data governance policy in WKC

In this data access policy use case example (Figure 6-2 on page 81), we use the classification capabilities of IBM Spectrum Discover to identify PII in the vast amounts of unstructured data, and then persist the PII that is tagged into the WKC. Then, data access policies within WKC are established to restrict access to any data that was tagged with PII.

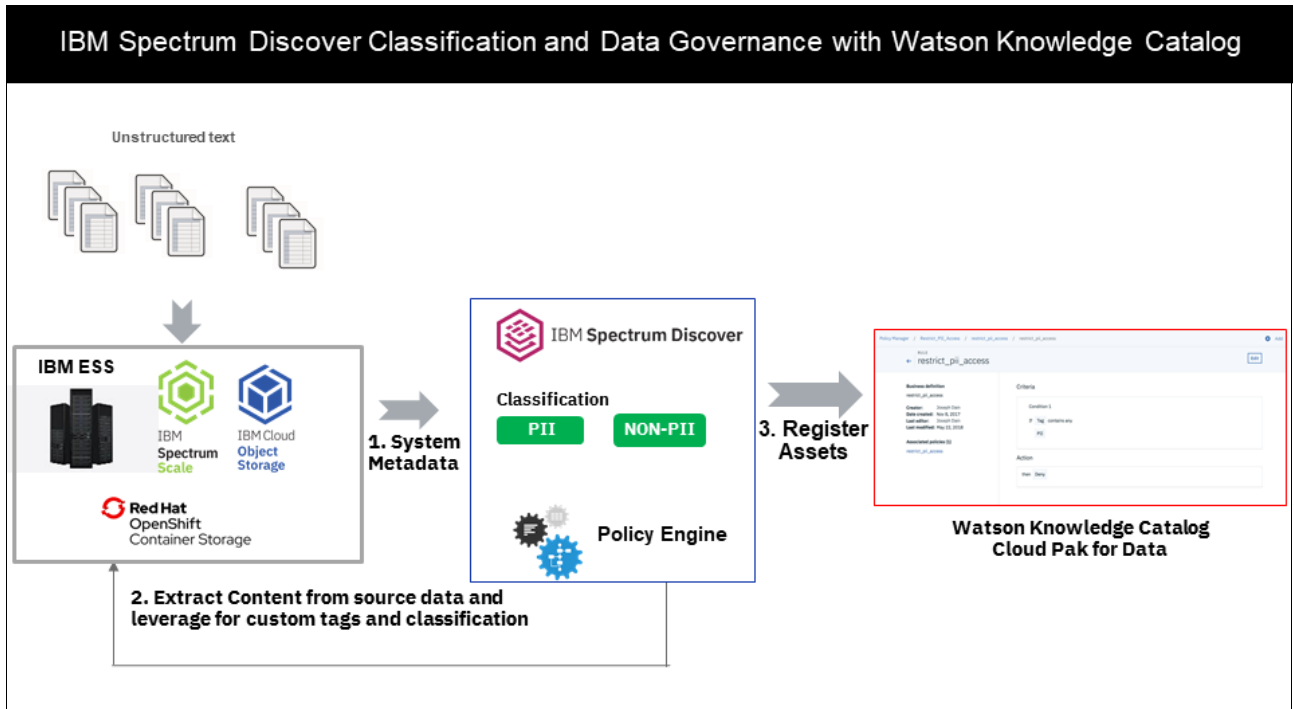


Figure 6-2 IBM Spectrum Discover classification capabilities to identify PII

In the IBM Spectrum Discover GUI, we use the research for a data governance data set that was classified by IBM Spectrum Discover. We want to select the PII that is tagged from the IBM Spectrum Discover GUI, as shown in Figure 6-3.

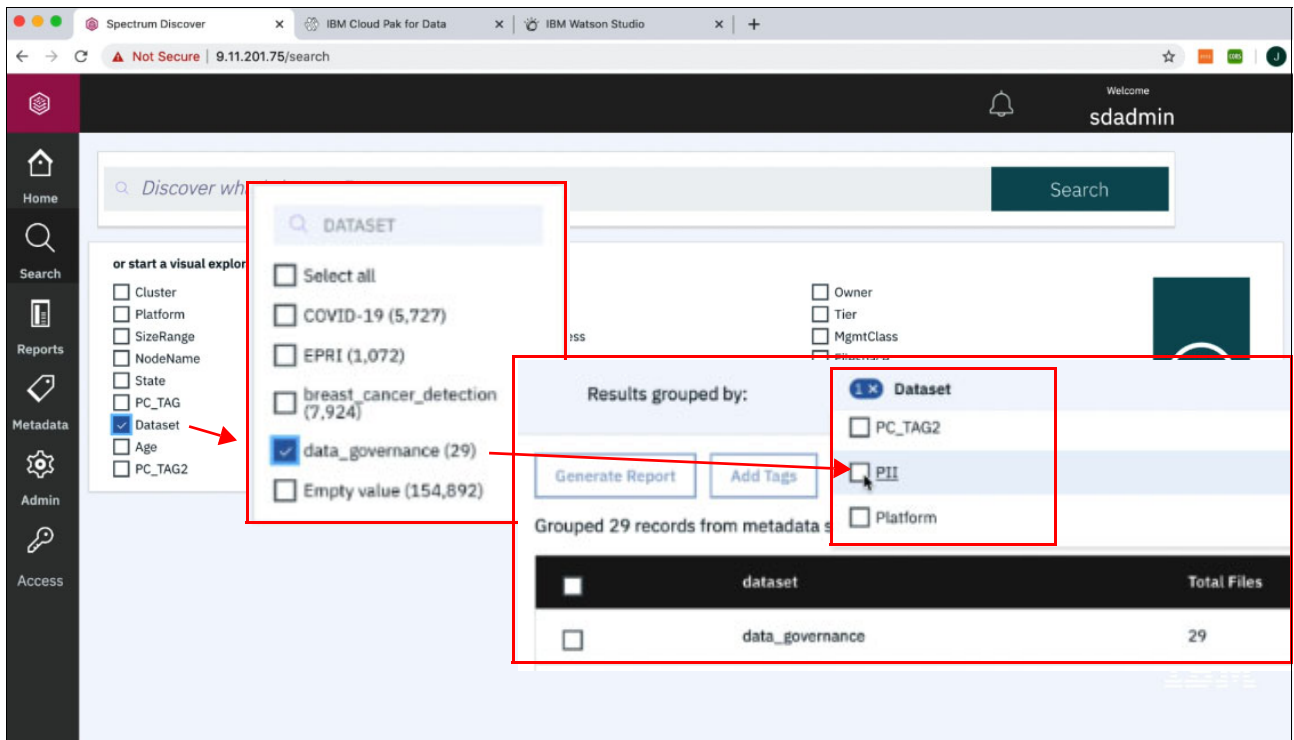


Figure 6-3 Selecting PII for a data governance use case

In this case (Figure 6-4), we can see that we have four files that contain PII.

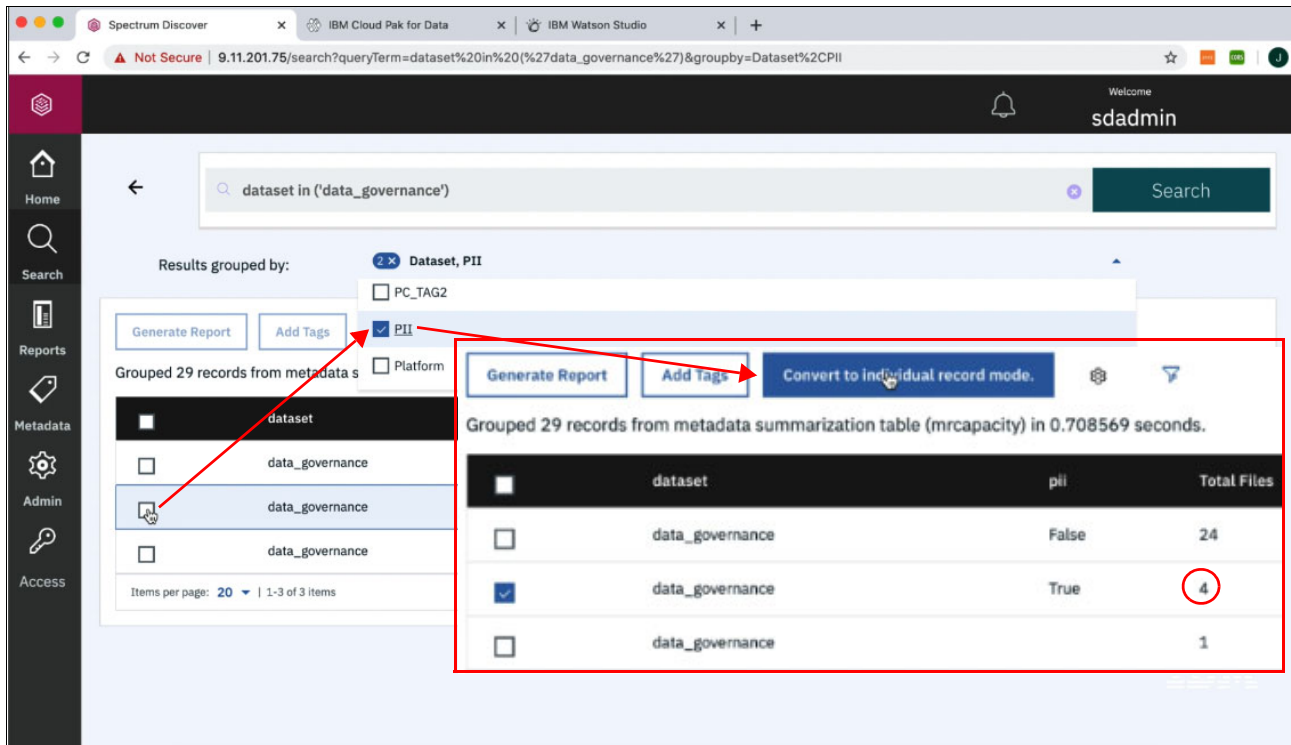


Figure 6-4 Converting PII for data governance to records

We want to select three of these files and export them to WKC, as shown in Figure 6-5.

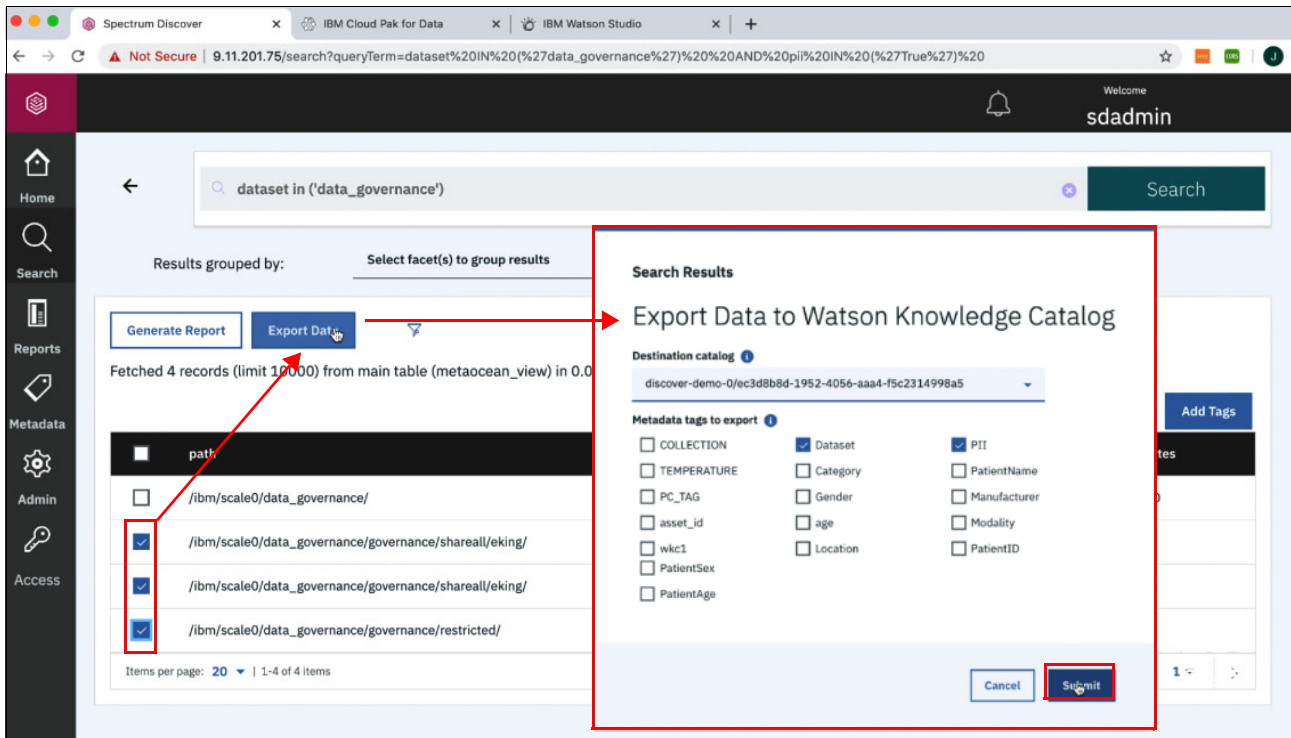


Figure 6-5 Selecting the records to export to Watson Knowledge Catalog

When exporting, we preserve the data set and PII tags. With WKC, we can see that the assets were added along with the PII and data set tags, as shown in Figure 6-6.

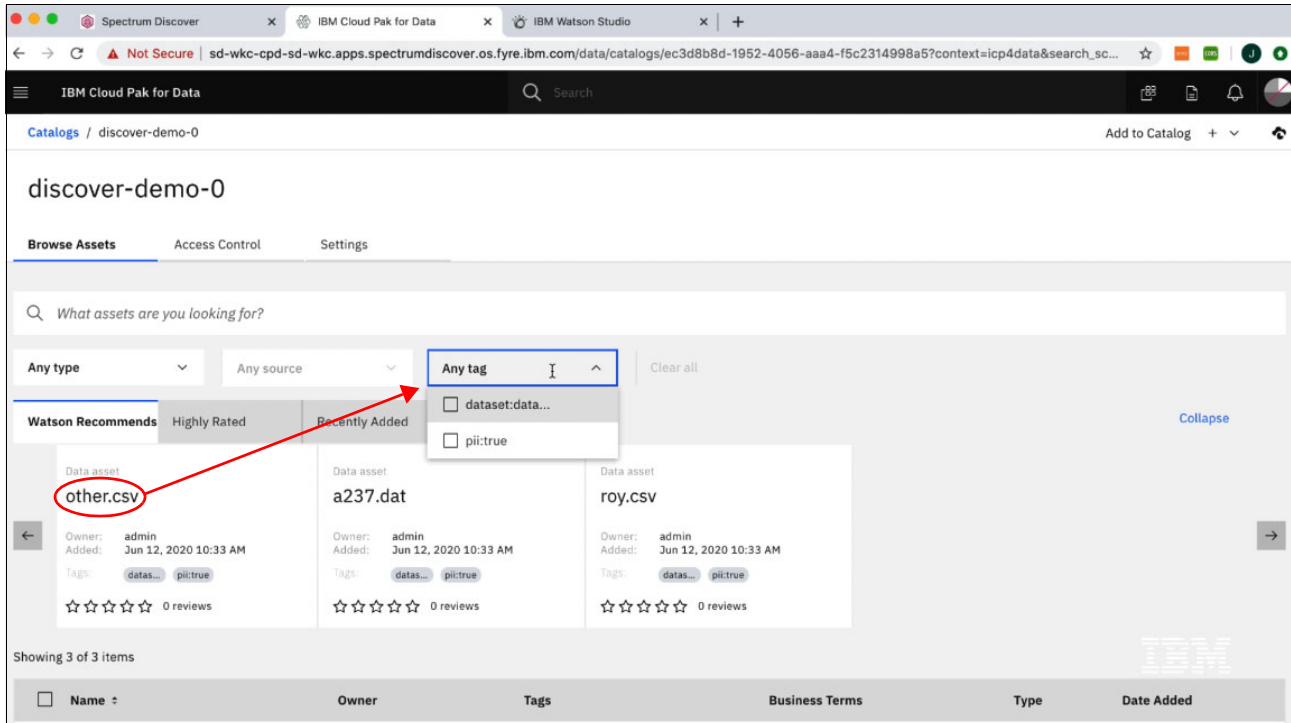


Figure 6-6 Selecting the other.csv record that was exported to Watson Knowledge Catalog

If we select and preview one of these assets, we can see that in column nine there is a credit card number (Figure 6-7) that was successfully identified by IBM Spectrum Discover.

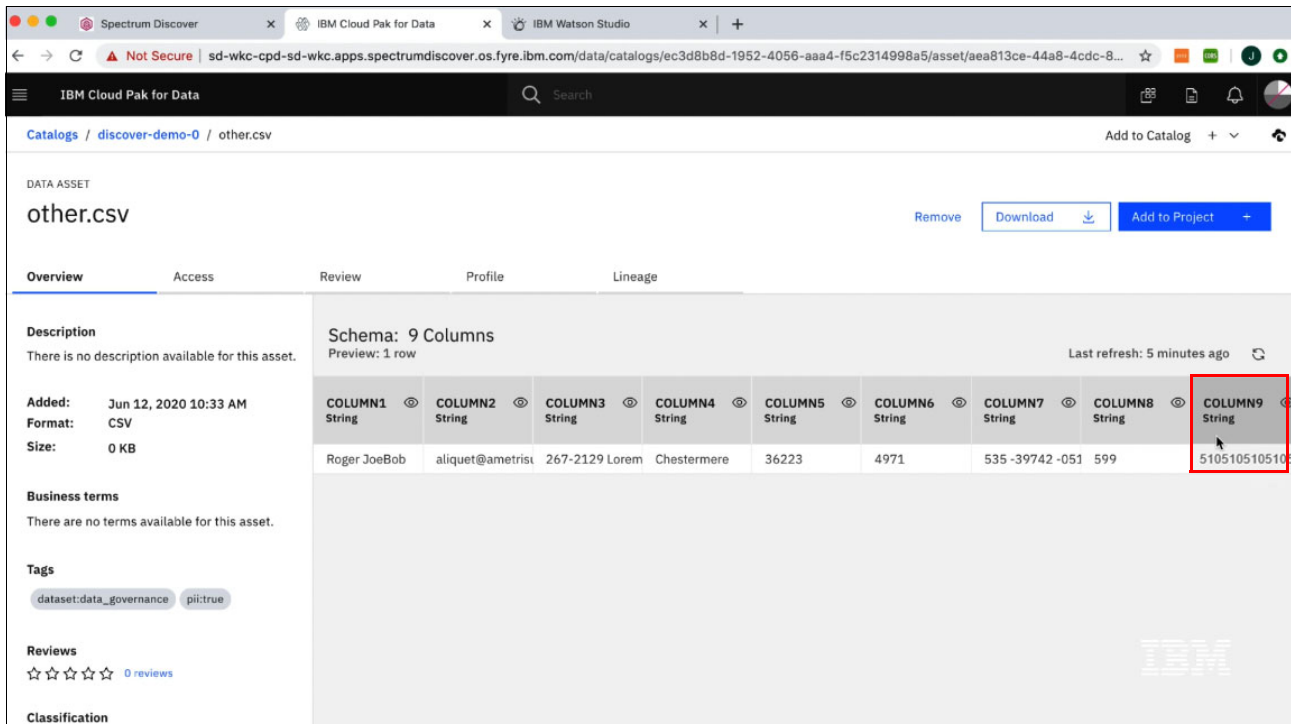


Figure 6-7 The asset contains a credit card number in column nine

Data governance policies and rules can be created in WKC to restrict access to any data that was tagged as PII by IBM Spectrum Discover, as shown in Figure 6-8.

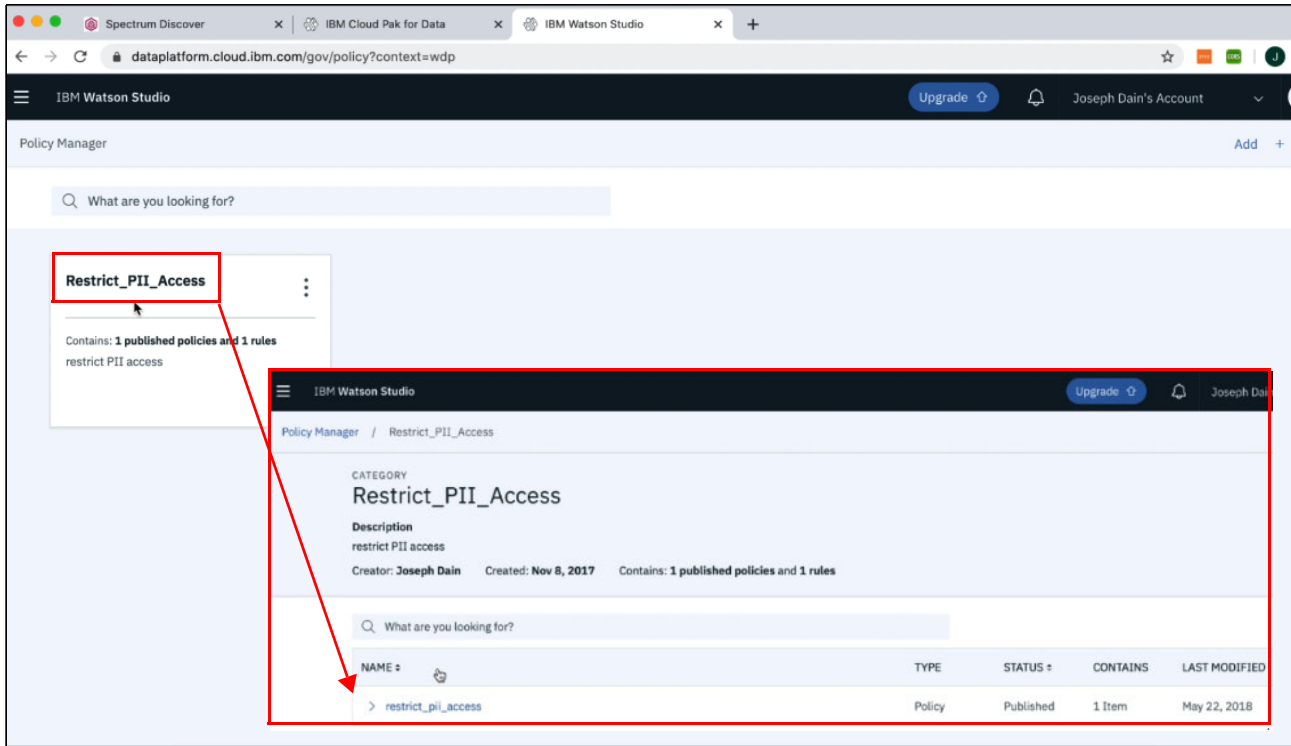


Figure 6-8 Showing the published policy and rule for restricting PII access

6.3.1 Creating a policy

Policies are used to describe and document your organization's guidelines, regulations, standards, or procedures to ensure that data and information assets are properly managed and used. Some examples of policies are Sensitive Data Handling and Data Sharing Agreement.

A policy is a natural-language description of a governance subject area. Policies describe how to control data. A policy consists of one or more rules. Each policy can:

- ▶ Contain multiple subpolicies.
- ▶ Reference a parent rule that must be sufficiently broad to encompass all of its subpolicies.
- ▶ Reference one or more governance rules to describe the characteristics for making information resources compliant with corporate objectives.
- ▶ Reference one or more data protection rules to create policies that specify types of data to restrict.
- ▶ Reference related artifacts, such as business terms and classifications.

You can organize policies in a hierarchy based on their meaning and relationships to each other.

By default, only data in relational data sets is protected. Through integration with IBM Spectrum Discover, data assets in unstructured data sources or any other types of assets that can be cataloged can be protected by policies.

All members of governed catalogs, regardless of their roles, are subject to policies. The owner of the data asset always sees the original values of that asset and is not subject to the policy restriction. For example:

- ▶ A policy that is named “High Quality Data” states that data must meet a high-quality standard.
- ▶ A subpolicy of “High Quality Data” called “High Quality Customer Data” states that customer data must meet a high-quality standard.
- ▶ The “High Quality Customer Data” policy references a governance rule called “Postal Codes Verification”.
- ▶ The “Postal Codes Verification” rule states that customer addresses must use valid postal codes, as provided by the post office.

All users can view published policies, but to author policies, you must have the “Manage governance artifacts” permission or be assigned as an author in the workflow that controls policies.

After the policy is created and published, the next step is to add rules for data protection.

Figure 6-9 shows various views of restrict_pii_access policy enforcement.

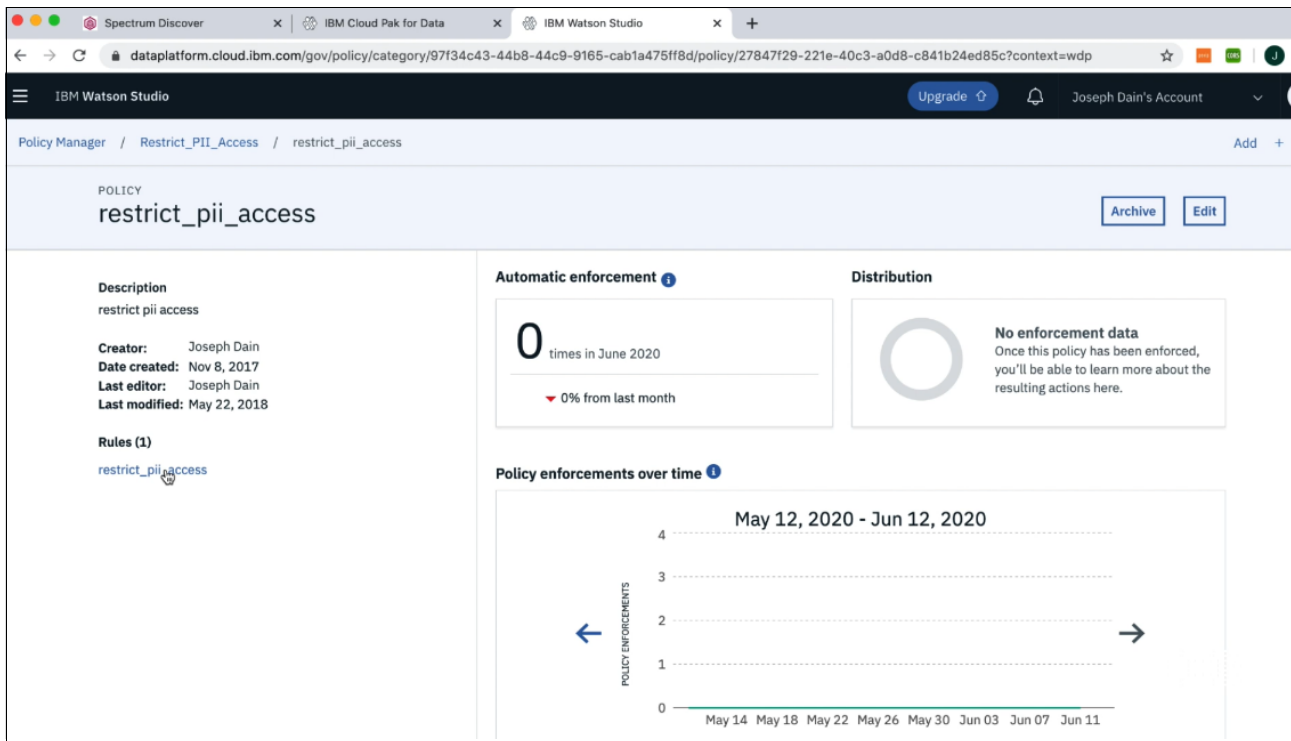


Figure 6-9 Restrict_pii_access policy enforcement

6.3.2 Creating rules for data protection

The policy must be published before you can add data protection rules to it.

Data protection rules control access to assets. Data protection rules are based on criteria, conditions, and an action that you define. They use predefined terms, such as business terms, data classes, or classifications, in expressions to define conditions.

You can define rules separately and later add them to published policies. They are artifacts that you can create, view, edit, rename, or delete. Rules that you created are added to the **Published** tab.

All users can view published data protection rules, but to author data protection rules, you must have the “Manage governance artifact” permission.

To create a rule, complete the following steps:

1. Select **Organize** → **Data and AI governance** → **Rules**, and then click **New rule**.
2. Select **Data protection rule** to mask sensitive data values or to deny access to data assets.
3. Specify the details and criteria for this rule:
 - a. The preselected policy type Access indicates that the purpose of this policy is to control the users' access to data.
 - b. Enter a business definition that explains what this rule does in plain language that is easy to understand. Include standard words and terms to make it easy to search for this rule.
4. Define the conditions in the rule builder:
 - a. The first term in a rule condition specifies an asset or user property. It can be one of the following items:
 - Asset owner: The email address of the user who owns the asset, for example, jblue@example.com.
 - Business term: The business term that was assigned to an asset or column, for example, Customer.
 - Data class: The classification of a column that categorizes the data, for example, Email Address.
 - Tag: The tag on an asset, for example, Marketing, Client Information, Claim.
 - User Name: The email address of a user requesting access to an asset, for example, jblue@example.com.
 - Classification: The type of sensitive information in the asset, for example, PII.
 - b. The operators of this condition depend on your selection for the first term. The operators must be the same within a condition and between condition blocks.
 - c. Depending on the first term of this condition, the second term can be one of the following items:
 - If there are values that are listed, choose a value from the list.
 - Type the name of a published artifact or select it from the list.

Note: You can select only published artifacts.

Otherwise, enter one or more values, such as tags, user IDs, or names:

- To enter a name or user ID, start typing the name or email address of a user.
 - To enter several tags manually, separate these tags with a comma.
- d. Specify the next conditions if required:
 - Click the plus-sign icon to specify more conditions.
 - Click the minus-sign icon to remove conditions.

- e. Select the action to take when the specified conditions are met:
 - Deny access to the asset.
 - Mask data.
 - f. Click **Create** to publish and activate the data protection rule. By default, data protection rules are enforced when they are in an active/published state, which means that no policy is required for the rule to be enforced in this case.
5. Click **Remove** if you want to delete the conditions that you defined for the Criteria property.
 6. Find or view published rules on the **Published** tab. Published rules are active and ready for use.

For each rule, you can edit its properties on the **Overview** tab.

7. Find or view published data protection rules on the **Related content** tab of the governance artifacts that are referenced or processed in data protection rules. You can then click a data protection rule to view its details or edit it.

In this example, which is shown in Figure 6-10, there is a restrict PII access rule that looks for any data that has a PII true tag set on it, and if this is found, it denies access to the data.

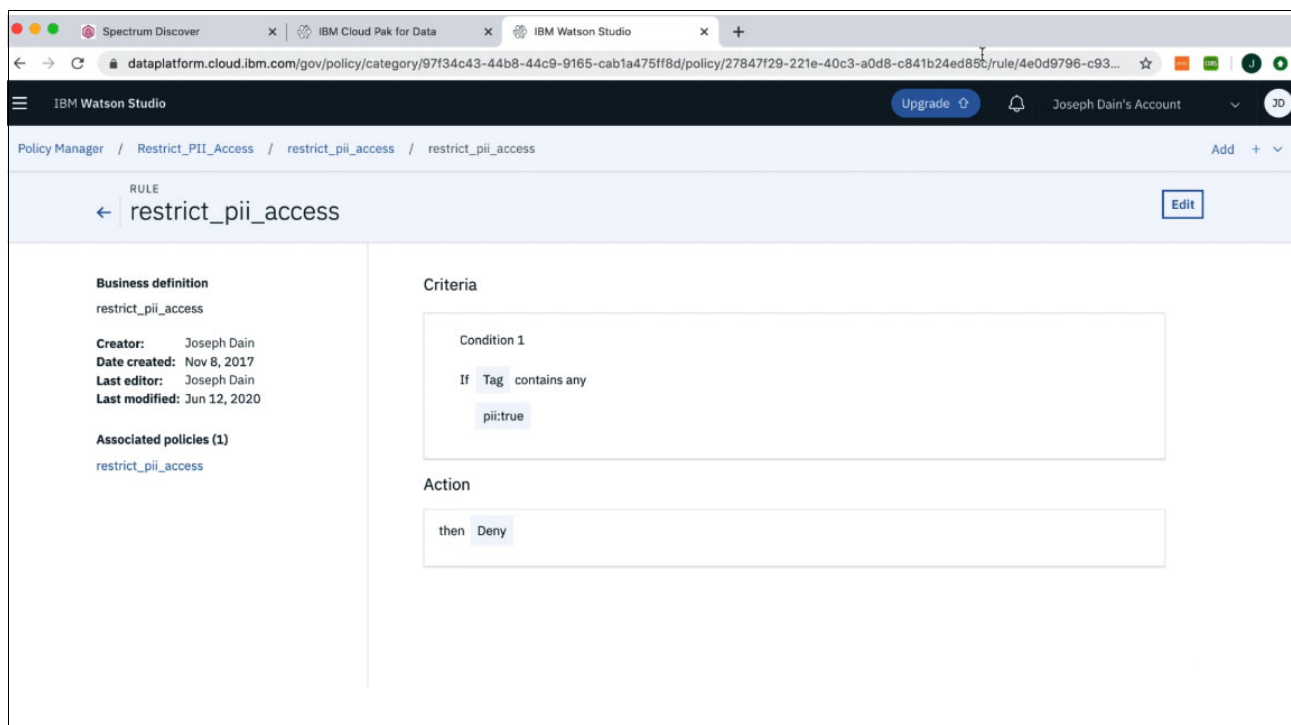


Figure 6-10 Setting to deny access if the PII tag is true



Conclusion

IBM Spectrum Discover is part of an overall portfolio for data and artificial intelligence (AI) solutions from IBM storage that helps customers gain visibility into their vast amounts of data and manage that data for AI workloads and analysis.

IBM Spectrum Discover and the IBM storage portfolio simplify solutions for AI workloads by providing an optimized infrastructure for each part of the AI ladder. The *AI ladder* is a prescriptive approach to accelerate your journey to AI. Data must be collected and organized for analysis. Finally, AI is infused throughout the organization.

Existing data management is not designed for massive amounts of data or to leverage data across multiple large data storage systems. Over 50% of companies say that they have data that is stored in too many silos, and everyone is struggling with capacity and modernization challenges. Business data is growing, and becoming more complex and difficult to manage.

Metadata is the key to creating structure around unstructured data. IBM Spectrum Discover creates a business-ready analytics foundation by building a state-of-the-art metadata catalog that unifies metadata for file and object storage wherever it is, whether it is on premises or in the cloud. In such an open and transparent data ecosystem, it is easy to discover, classify, label, and find and activate data for large-scale analytics and data science.

Using IBM Spectrum Discover, users can automate cataloging or indexing for easier collaboration. They can find and identify relevant data faster, ensure that real-time data updates are current and integrated, and enable comprehensive insight, data governance, and data optimization.

The insights from IBM Spectrum Discover can be integrated with IBM Watson Knowledge Catalog (WKC) in IBM Cloud Pak for Data (IBM CP4D) to provide an AI and analytics solution that seamlessly ingests and uses data from IBM Cloud Object Storage (IBM COS), IBM Spectrum Scale, IBM Elastic Storage Systems, Red Hat Ceph, Amazon Web Services (AWS) Simple Storage Service (S3) storage, Isilon, NetApp, and Server Message Block (SMB) storage for analysis by IBM Watson Solutions. WKC provides capabilities for AI analysis that have not been available, and unlocks data from medical, financial, media, manufacturing, retail, and many more business opportunities that can leverage AI.

Integration of IBM Spectrum Discover with WKC and IBM CP4D provides improved data quality for increased productivity, provides image video indexing and tagging for faster analysis and search by using advanced AI with IBM Watson solutions and IBM CP4D, enables efficient organization and preparation of vast amounts of data for AI workflows, and identifies personal data for anomalies and compliance.

Related publications

The publications that are listed in this section are considered suitable for a more detailed description of the topics that are covered in this paper.

IBM Redbooks

The following IBM Redbooks publication provides more information about the topics in this document. It might be available in softcopy only.

- ▶ *IBM Spectrum Discover: Metadata Management for Deep Insight of Unstructured Storage*, REDP-5550

You can search for, view, download, or order these documents and other Redbooks, Redpapers, web docs, drafts, and additional materials, at the following website:

ibm.com/redbooks

Other publications

This publication is also relevant as a further information source:

- ▶ Wadleigh, et al, *Software Optimization for High Performance Computing: Creating Faster Applications 1st Edition*, Prentice Hall, 2000, ISBN 0130170089

Online resources

These websites are also relevant as further information sources:

- ▶ IBM Cloud Pak for Data: IBM Watson Knowledge Catalog
https://www.ibm.com/support/producthub/icpdata/docs/content/SSQNUZ_current/ws/catalog/overview-wkc.html
- ▶ IBM Spectrum Discover Free 90 Day Trial
<https://www.ibm.com/products/spectrum-discover>
- ▶ IBM Spectrum Discover (IBM Knowledge Center)
<https://www.ibm.com/support/knowledgecenter/SSY8AC>
- ▶ IBM Spectrum Discover: Integrate images for easier analysis to the AI Journey (COVID-19 Use Case)
<https://youtu.be/RWawwEvCfkE>
- ▶ IBM Spectrum Discover Video Series Overview: Faster Deployment to Gain Insights Sooner
<https://www.youtube.com/watch?v=LtMhVzI2SEQ>
- ▶ IBM Visual Insights
<https://www.ibm.com/products/ibm-visual-insights>

- ▶ IBM Visual Insights Developer Portal (Overview, Features, Trial, and Resources)
<https://developer.ibm.com/linuxonpower/deep-learning-powerai/vision>
- ▶ IBM Visual Insights (IBM Knowledge Center)
<https://www.ibm.com/support/knowledgecenter/SSRU69>
- ▶ IBM Visual Insights and IBM Spectrum Discover: Automation with Computer Vision (COVID-19 Use Case)
<https://www.youtube.com/watch?v=rVuki05vcHs>
- ▶ IBM Watson Knowledge Catalog
<https://www.ibm.com/cloud/watson-knowledge-catalog>
- ▶ IBM Watson Knowledge Catalog overview (IBM Knowledge Center)
https://www.ibm.com/support/knowledgecenter/en/SSZJPZ_11.7.0/wsj/catalog/overview-wkc.html

GitHub resources

- ▶ Dockerhub example
<https://hub.docker.com/r/ibmcom/spectrum-discover-example-application>
- ▶ IBM Spectrum Discover Application Catalog
https://github.com/IBM/Spectrum_Discover_App_Catalog
- ▶ IBM Spectrum Discover Application SDK
https://github.com/IBM/Spectrum_Discover_Application_SDK

visJupyter display library resource

- ▶ University of California San Diego School of Medicine visJS2Jupyter display library
<https://ucsd-ccbb.github.io/visJS2jupyter>

Help from IBM

IBM Support and downloads

[ibm.com/support](https://www.ibm.com/support)

IBM Global Services

[ibm.com/services](https://www.ibm.com/services)



REDP-5603-00

ISBN 073845902x

Printed in U.S.A.

Get connected

