Comparison of 2 Data Catalogues Tools

presented by Lukas Heubach and Leonhard Krause
DWH-TINF18D

Agenda

- What is a Data Catalog?
- Why Data Catalogues?
- IBM InfoSphere IGC
- Lumada Data Catalog
- Comparison
- Sources

What is a Data Catalog?

- digital inventory (directory)
- contains all company data
- Single source of trust - data inventory that is correct and can be relied upon
- Data catalog is filled with metadata of technical and business origin
- Data supply & demand
- provides functions for registering, retrieving, using, evaluating and analyzing data

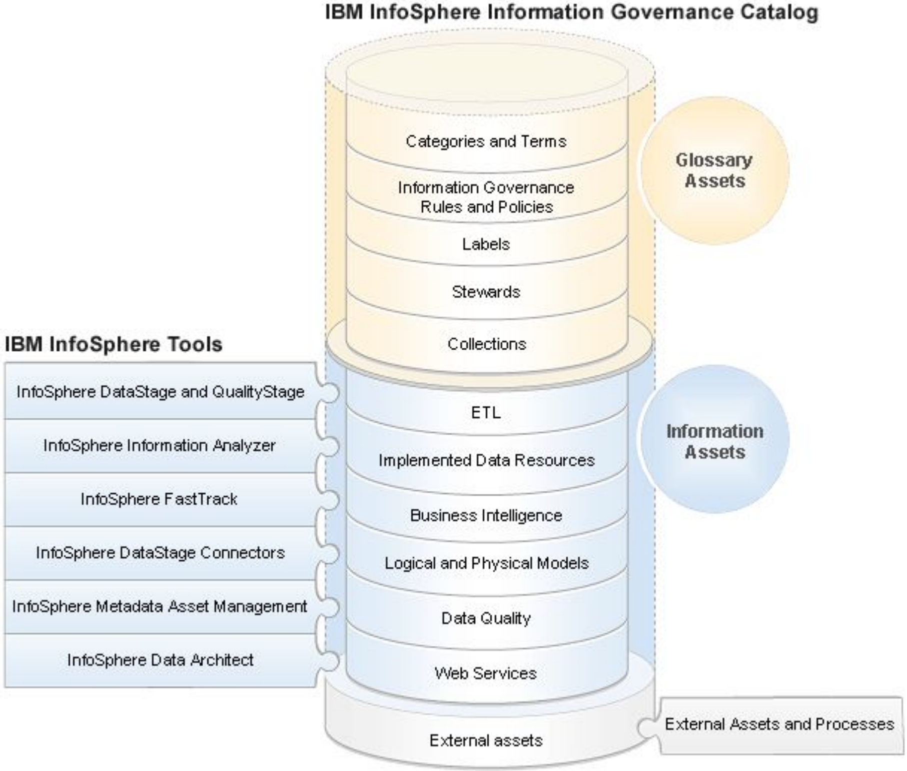
Why Data Catalogues?

- Data is constantly accumulating, more and more and in new formats
- data sets should be transparently available in the company
- to organize company data
- Main objective: to promote collaboration within the company by making relevant data

IBM InfoSphere IGC – General

InfoSphere Information Governance Catalog

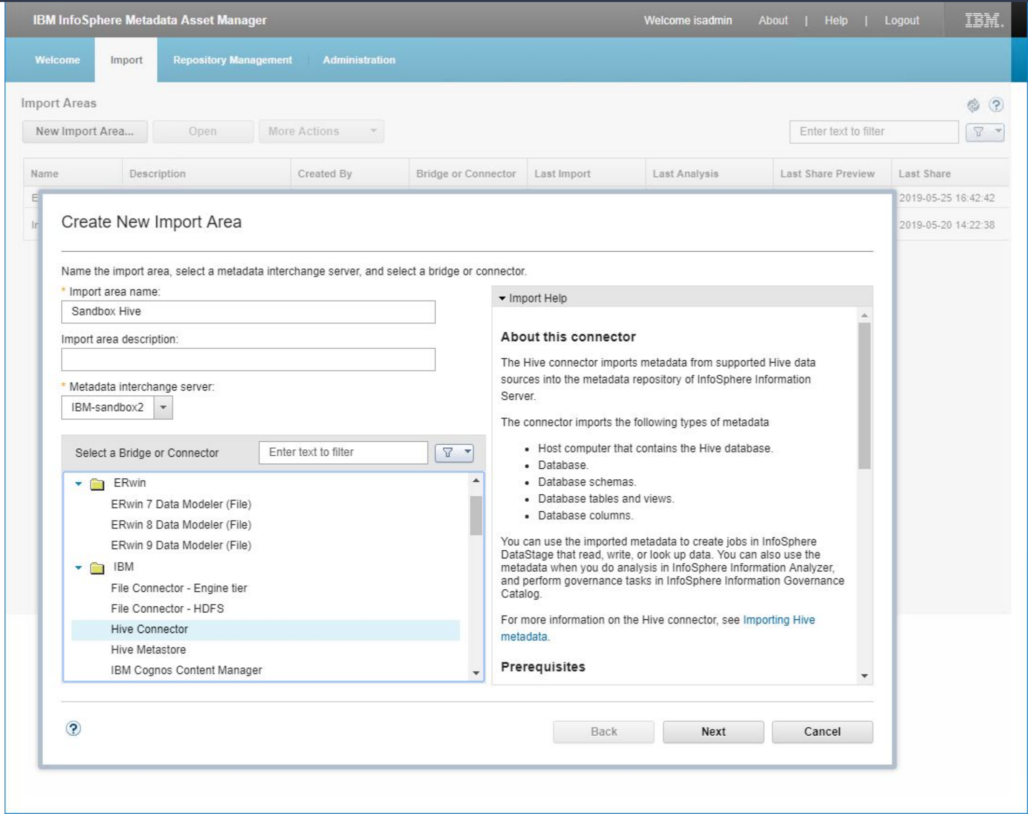
- web-based tool
- Create, manage, share, use business knowledge
- Prices
- individual offer
- Goals
  - Data -> reliable information
- Can be used in conjunction with other InfoSphere tools


IBM InfoSphere Information Governance Catalog
Categories and Terms
Information Governance Rules and Policies
Labels
Stewards
Collections
Glossary Assets
IBM InfoSphere Tools
InfoSphere DataStage and QualityStage
InfoSphere Information Analyzer
InfoSphere FastTrack
InfoSphere DataStage Connectors
InfoSphere Metadata Asset Management
InfoSphere Data Architect
ETL
Implemented Data Resources
Business Intelligence
Logical and Physical Models
Data Quality
Web Services
Information Assets
External assets
External Assets and Processes

IBM InfoSphere IGC – Functions

Connection of data sources

- different types of sources (asset types)
    - (AWS S3, IBM InfoSphere DB2, Oracle ...)
- Import e.g. via Metadata Asset Manager



IBM InfoSphere Metadata Asset Manager — Welcome isadmin | About | Help | Logout

Welcome | Import | Repository Management | Administration

Import Areas

New Import Area... | Open | More Actions | Enter text to filter

Name	Description	Created By	Bridge or Connector	Last Import	Last Analysis	Last Share Preview	Last Share
E							2019-05-25 16:42:42
Ir							2019-05-20 14:22:38

Create New Import Area

Name the import area, select a metadata interchange server, and select a bridge or connector.

* Import area name: Sandbox Hive

Import area description:

* Metadata interchange server: IBM-sandbox2

Select a Bridge or Connector — Enter text to filter

- ERwin
    - ERwin 7 Data Modeler (File)
    - ERwin 8 Data Modeler (File)
    - ERwin 9 Data Modeler (File)
- IBM
    - File Connector - Engine tier
    - File Connector - HDFS
    - Hive Connector
    - Hive Metastore
    - IBM Cognos Content Manager

Import Help

About this connector

The Hive connector imports metadata from supported Hive data sources into the metadata repository of InfoSphere Information Server.

The connector imports the following types of metadata

- Host computer that contains the Hive database.
- Database.
- Database schemas.
- Database tables and views.
- Database columns.

You can use the imported metadata to create jobs in InfoSphere DataStage that read, write, or look up data. You can also use the metadata when you do analysis in InfoSphere Information Analyzer, and perform governance tasks in InfoSphere Information Governance Catalog.

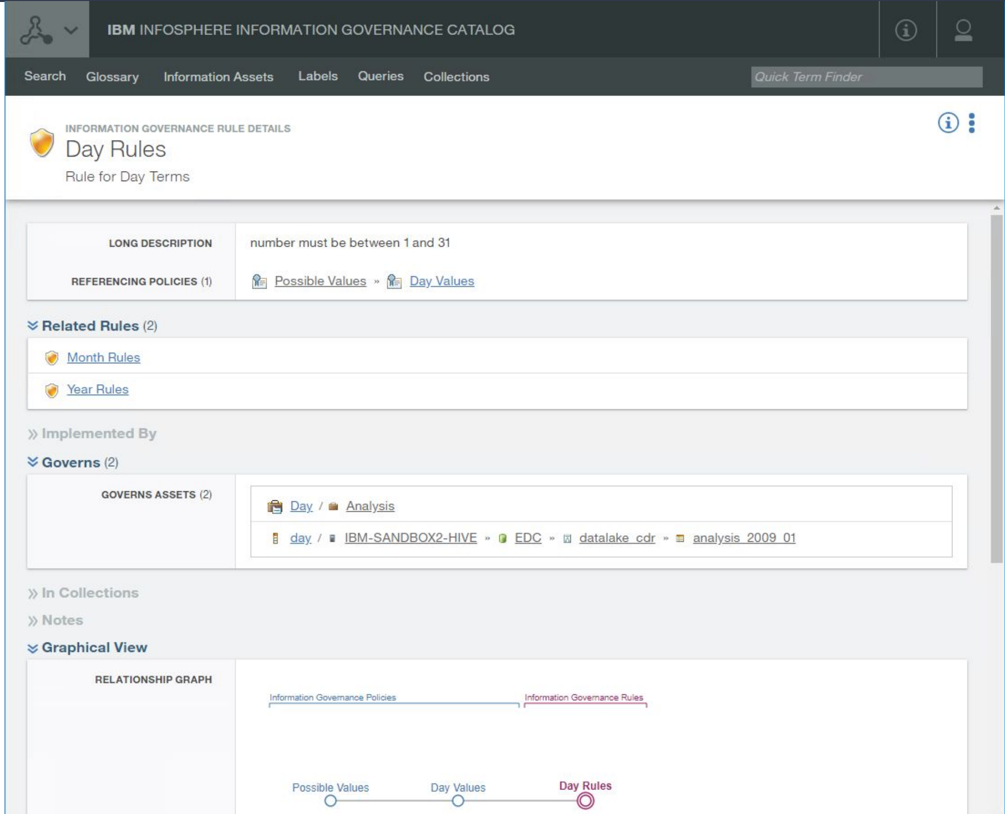
For more information on the Hive connector, see Importing Hive metadata.

Prerequisites

IBM InfoSphere IGC – Functions

Glossary Assets

- create/represent complex relationships between assets
- Categories
    - like a folder to structure Glossary Assets
- Terms
    - Word/phrase that describes a characteristic
- IG Rules
    - a natural language description of a criterion that determines whether an information asset meets a business objective.
- IG Policies
    - a natural language description of a subject area
- Labels



INFORMATION GOVERNANCE RULE DETAILS

Day Rules

Rule for Day Terms

LONG DESCRIPTION: number must be between 1 and 31

REFERENCING POLICIES (1): Possible Values » Day Values

Related Rules (2)

- Month Rules
- Year Rules

Governs (2)

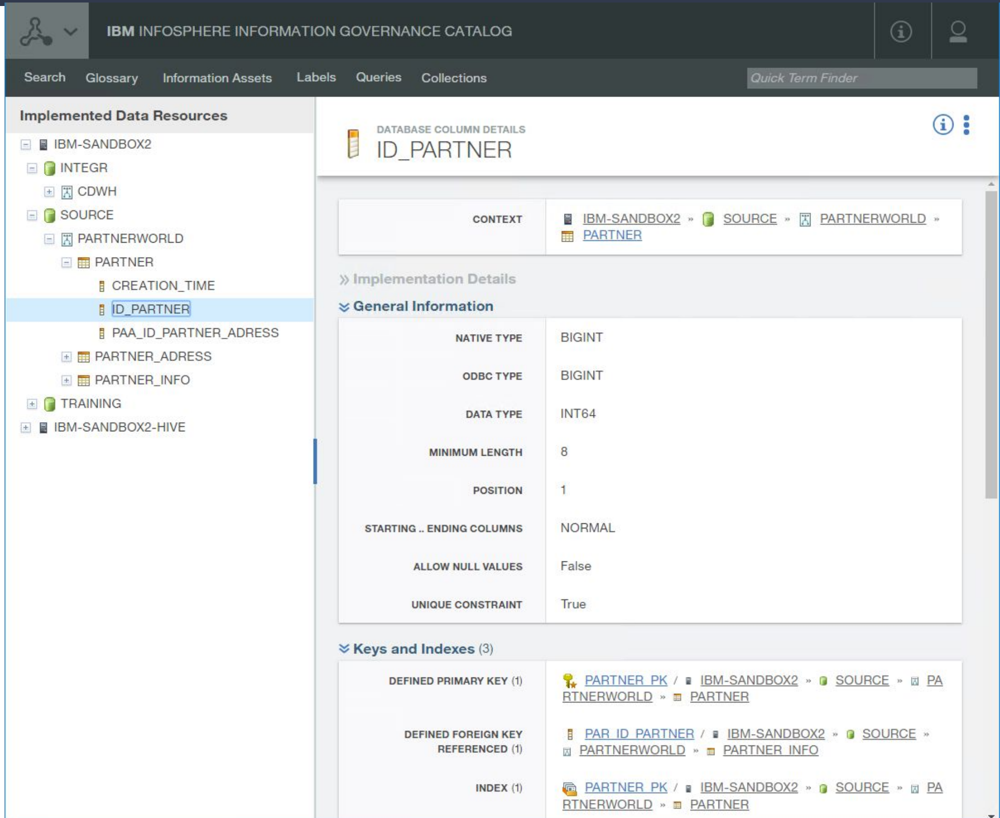
GOVERNS ASSETS (2):

- Day / Analysis
- day / IBM-SANDBOX2-HIVE » EDC » datalake_cdr » analysis_2009_01

IBM InfoSphere IGC – Functions

Information Assets

- Imported records
- Imported metadata
- Display of all included data sources
- Display of all data sets/metadata
- Assign to Glossary Assets



Implemented Data Resources

- IBM-SANDBOX2
  - INTEGR
    - CDWH
  - SOURCE
    - PARTNERWORLD
      - PARTNER
        - CREATION_TIME
        - ID_PARTNER
        - PAA_ID_PARTNER_ADRESS
      - PARTNER_ADRESS
      - PARTNER_INFO
  - TRAINING
- IBM-SANDBOX2-HIVE

DATABASE COLUMN DETAILS

ID_PARTNER

CONTEXT	IBM-SANDBOX2 » SOURCE » PARTNERWORLD » PARTNER

» Implementation Details

General Information

NATIVE TYPE	BIGINT
ODBC TYPE	BIGINT
DATA TYPE	INT64
MINIMUM LENGTH	8
POSITION	1
STARTING .. ENDING COLUMNS	NORMAL
ALLOW NULL VALUES	False
UNIQUE CONSTRAINT	True

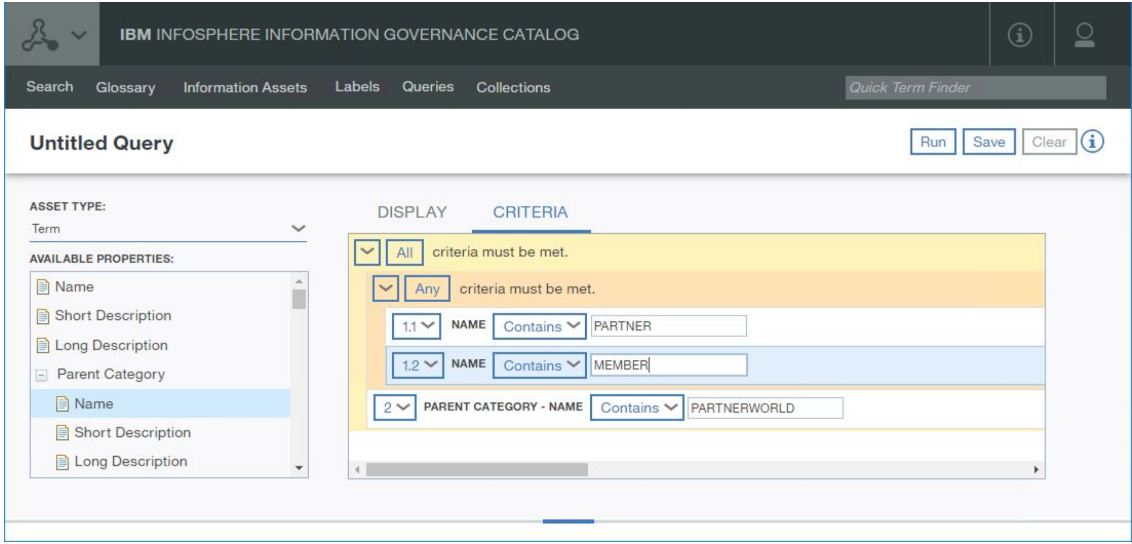
Keys and Indexes (3)

DEFINED PRIMARY KEY (1)	PARTNER_PK / IBM-SANDBOX2 » SOURCE » PARTNERWORLD » PARTNER
DEFINED FOREIGN KEY REFERENCED (1)	PAR_ID_PARTNER / IBM-SANDBOX2 » SOURCE » PARTNERWORLD » PARTNER_INFO
INDEX (1)	PARTNER_PK / IBM-SANDBOX2 » SOURCE » PARTNERWORLD » PARTNER

IBM InfoSphere IGC – Functions

Queries

- create your own queries
- for Information Assets
- for Glossary Assets
- result: Table with information



Untitled Query

ASSET TYPE: Term

All criteria must be met.

Any criteria must be met.

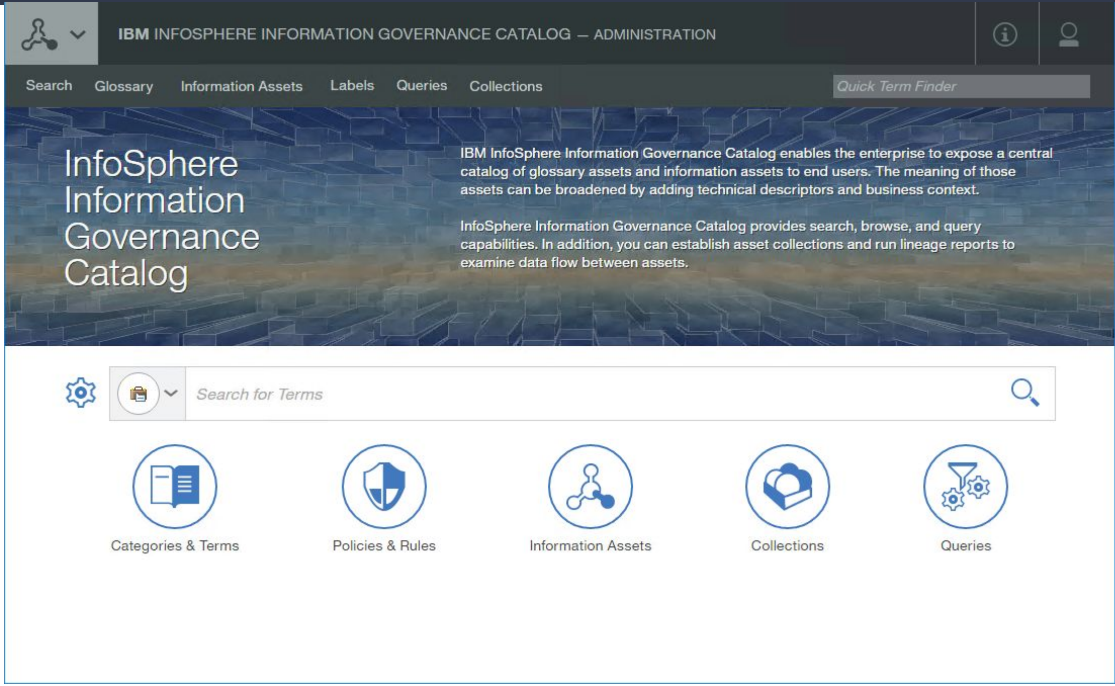
1.1 NAME Contains PARTNER

1.2 NAME Contains MEMBER

2 PARENT CATEGORY - NAME Contains PARTNERWORLD

IBM InfoSphere IGC – Who is it for?

- Business Analysts
- Business experts
- Organizations that
  - want to manage a common enterprise vocabulary and governance practices
  - want to leverage the potential of integrated metadata
  - reduce the need for technical training



IBM INFOSPHERE INFORMATION GOVERNANCE CATALOG — ADMINISTRATION

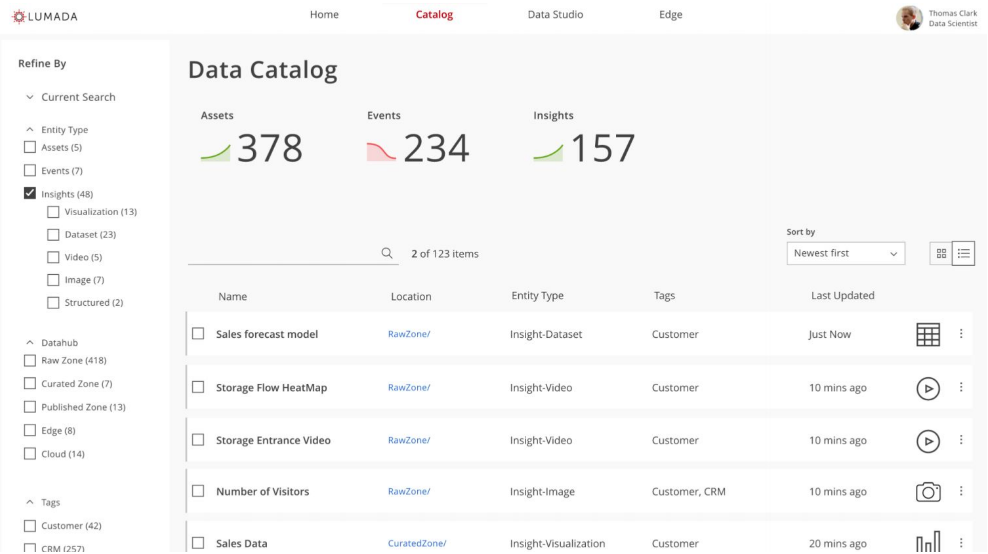
InfoSphere Information Governance Catalog

IBM InfoSphere Information Governance Catalog enables the enterprise to expose a central catalog of glossary assets and information assets to end users. The meaning of those assets can be broadened by adding technical descriptors and business context.

InfoSphere Information Governance Catalog provides search, browse, and query capabilities. In addition, you can establish asset collections and run lineage reports to examine data flow between assets.

Lumada Data Catalog (Waterline Data Catalog)

- Tool for managing data from diverse sources
- uses machine learning to build data inventory
- patented fingerprinting technology for data
- Management of data lakes
- Goal: analyze large amounts of data automatically
- Price and demo are only available on request



Refine By

Current Search

Entity Type
□ Assets (5)
□ Events (7)
☑ Insights (48)
□ Visualization (13)
□ Dataset (23)
□ Video (5)
□ Image (7)
□ Structured (2)

Datahub
□ Raw Zone (418)
□ Curated Zone (7)
□ Published Zone (13)
□ Edge (8)
□ Cloud (14)

Tags
□ Customer (42)
□ CRM (257)

Data Catalog

Assets	Events	Insights
378	234	157

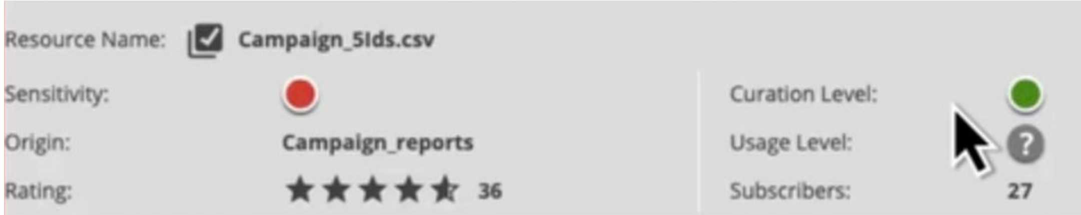
2 of 123 items

Sort by: Newest first

	Name	Location	Entity Type	Tags	Last Updated
□	Sales forecast model	RawZone/	Insight-Dataset	Customer	Just Now
□	Storage Flow HeatMap	RawZone/	Insight-Video	Customer	10 mins ago
□	Storage Entrance Video	RawZone/	Insight-Video	Customer	10 mins ago
□	Number of Visitors	RawZone/	Insight-Image	Customer, CRM	10 mins ago
□	Sales Data	CuratedZone/	Insight-Visualization	Customer	20 mins ago

Lumada Data Catalog – Functions

- Data recognition using metadata-based search
- Management of sensitive data
- Patented data fingerprinting
  - data processing based on machine learning
  - automatic recognition of sensitive data
  - enables Google-like search with corporate identifiers
- visualization of data, -origins and -relationships
- Detection of redundant data
- comment, subscription, rating function
- Recording of user data

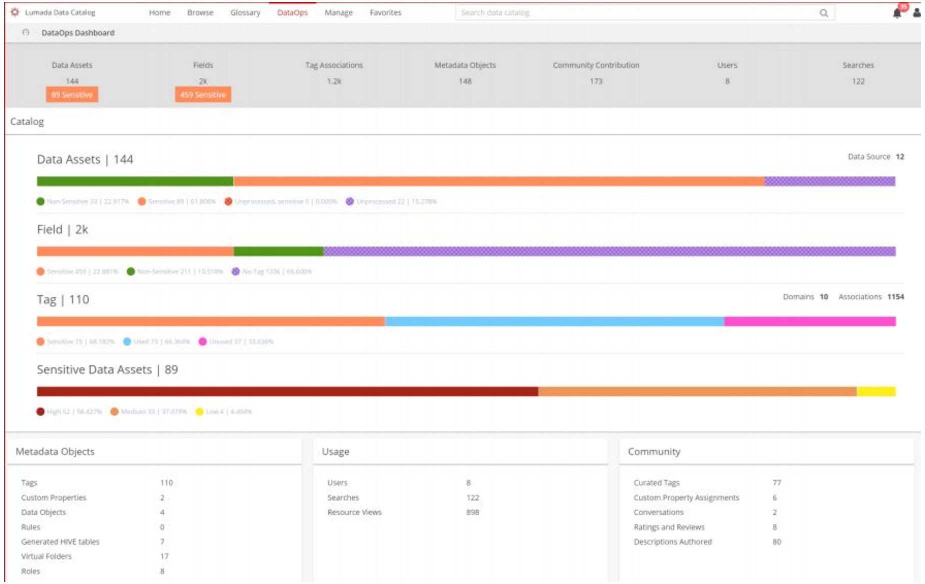


Resource Name: ☑ Campaign_5Ids.csv

Sensitivity:	🔴	Curation Level:	🟢
Origin:	Campaign_reports	Usage Level:	?
Rating:	★★★★⯪ 36	Subscribers:	27

Lumada Data Catalog – Who is it for?

- Companies with large amounts of data
- Data analysts
- Fast and efficient compliance with data protection regulations
- Management of sensitive data



DataOps Dashboard

Data Assets	Fields	Tag Associations	Metadata Objects	Community Contribution	Users	Searches
144 89 Sensitive	2k 459 Sensitive	1.2k	148	173	8	122

Catalog

Metadata Objects

Tags	110
Custom Properties	2
Data Objects	4
Rules	0
Generated HIVE tables	7
Virtual Folders	17
Roles	8

Usage

Users	8
Searches	122
Resource Views	898

Community

Curated Tags	77
Custom Property Assignments	6
Conversations	2
Ratings and Reviews	8
Descriptions Authored	80

Comparison

Category	**IBM InfoSphere IGC**	**Lumada Data Catalogue**
Similarities	● Data Catalog Tools ● individual price offer ● use of different data sources	
User Interface	Web Tool	no specification (possibly desktop application)
Focus	data management	data analysis
AI	No	AI-based data analysis and structuring
Product Portfolio	IBM InfoSphere Family	Lumada Data Services
“Social-Media” Tools	No	Yes (comments, subscription, rating)

Sources

- https://www.hitachivantara.com/de-de/products/data-management-analytics/lumada-data-services/lumada-data-catalog.html (retrieved on: 13.02.2021 at 14:32 Uhr)
- https://www.hitachivantara.com/en-us/products/data-management-analytics/lumada-data-catalog.html (retrieved on: 13.02.2021 at 14:32 Uhr)
- https://www.hitachivantara.com/en-us/pdfd/datasheet/lumada-data-lake-datasheet.pdf (abgerufen am: 13.02.2021 at 14:32 Uhr)
- https://www.hitachivantara.com/de-de/pdf/datasheet/lumada-data-catalog-datasheet-de.pdf (retrieved on: 13.02.2021 at 14:32 Uhr)
- https://www.talend.com/de/resources/what-is-data-catalog/(retrieved on: 13.02.2021 at 14:32 Uhr)
- https://www.cc-cdq.ch/data-catalogs (retrieved on: 13.02.2021 at 14:32 Uhr)
- https://www.ibm.com/de-de/marketplace/information-governance-catalog (retrieved on: 13.02.2021 at 14:32 Uhr)
- https://www.saracus.com/blog/der-ibm-infosphere-information-governance-catalog/ (retrieved on: 13.02.2021 at 14:32 Uhr)

THANK YOU FOR YOUR ATTENTION!