


Vergleich von 2 Data Catalogues Tools

präsentiert von Lukas Heubach und Leonhard Krause
DWH-TINF18D

A dark blue diagonal gradient bar that starts from the bottom left and extends towards the top right, covering the lower half of the slide.

Agenda

- Was ist ein Data Catalog?
- Wozu wird es benötigt?
- IBM InfoSphere IGC
- Lumada Data Catalog
- Vergleich
- Quellen

Was ist ein Data Catalog?

- digitales Inventar (Verzeichnis)
- enthält sämtliche Unternehmensdaten
- Single Source of Trust - Datenbestand der korrekt ist und man kann sich auf diesen verlassen
- Datenkatalog wird mit Metadaten technischer und fachlicher Herkunft befüllt
- Datenangebot & -nachfrage
- bietet Funktionen zum Registrieren, Abrufen, Verwenden, Bewerten und Analysieren von Daten

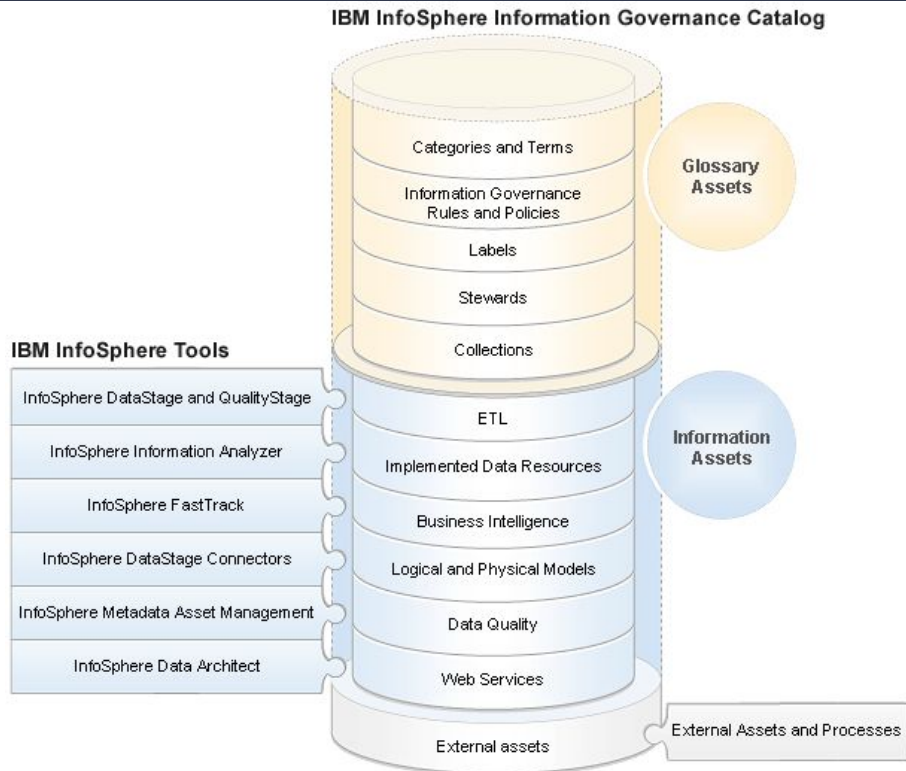
Wozu Data Catalogues?

- ständig fallen Daten an, immer mehr und in neuen Formaten
- Datensätze sollen im Unternehmen transparent verfügbar sein
- ordnen von Unternehmensdaten
- Hauptziel: Zusammenarbeit im Unternehmen fördern, durch Verfügbarkeit von relevanten Daten (einheitlich katalogisiert)

IBM InfoSphere IGC – Allgemein

InfoSphere Information Governance Catalog

- webbasiertes Tool
- Erzeugen, Verwalten, Teilen, Nutzen von Geschäftswissen
- Preise
 - individuelles Angebot
- Ziele
 - Daten -> zuverlässige Informationen
- In Verbindung mit anderen InfoSphere Tools nutzbar



IBM InfoSphere IGC – Funktionen

Anbindung von Datenquellen

- verschiedene Quellenarten (Asset Types)
 - (AWS S3, IBM InfoSphere DB2, Oracle ...)
- Import z.B. über Metadata Asset Manager

The screenshot shows the 'Create New Import Area' dialog in the IBM InfoSphere Metadata Asset Manager. The dialog is titled 'Create New Import Area' and contains the following fields and sections:

- Name the import area, select a metadata interchange server, and select a bridge or connector.**
- Import area name:** A text field containing 'Sandbox Hive'.
- Import area description:** An empty text field.
- Metadata interchange server:** A dropdown menu showing 'IBM-sandbox2'.
- Select a Bridge or Connector:** A list box showing a tree structure with 'ERwin' and 'IBM' folders. Under 'IBM', 'Hive Connector' is selected.
- Import Help:** A section titled 'About this connector' providing information about the Hive connector, including the types of metadata it imports (Host computer, Database, Database schemas, Database tables and views, Database columns) and a link to 'Importing Hive metadata'.
- Prerequisites:** A section for prerequisites, currently empty.
- Buttons:** 'Back', 'Next', and 'Cancel' buttons at the bottom right.

IBM InfoSphere IGC – Funktionen

Glossary Assets

- Anlegen/Darstellen von komplexen Beziehungen zwischen Assets
- Categories
 - Wie ein Ordner, um Glossary Assets zu strukturieren
- Terms
 - Wort/Phrase, die eine Charakteristik beschreibt
- IG Rules
 - Eine Beschreibung eines Kriteriums in natürlicher Sprache, das bestimmt, ob ein Information Asset einem Business Objective entspricht
- IG Policies
 - Eine Beschreibung in natürlicher Sprache eines Themengebiets
- Labels

The screenshot displays the IBM InfoSphere Information Governance Catalog interface. The top navigation bar includes 'Search', 'Glossary', 'Information Assets', 'Labels', 'Queries', and 'Collections'. The main content area is titled 'INFORMATION GOVERNANCE RULE DETAILS' and shows the 'Day Rules' rule for 'Rule for Day Terms'. The rule details include a 'LONG DESCRIPTION' (number must be between 1 and 31) and 'REFERENCING POLICIES (1)' (Possible Values, Day Values). Below this, there are sections for 'Related Rules (2)' (Month Rules, Year Rules), 'Implemented By', 'Governs (2)' (Day / Analysis, IBM-SANDBOX2-HIVE, EDC, datalake cdr, analysis_2009_01), 'In Collections', 'Notes', and 'Graphical View'. The 'Graphical View' section shows a 'RELATIONSHIP GRAPH' with nodes for 'Possible Values', 'Day Values', and 'Day Rules' connected by lines.

IBM InfoSphere IGC – Funktionen

Information Assets

- Importierte Datensätze
- Importierte Metadaten
- Anzeige aller eingebundenen Datenquellen
- Anzeige aller Datensätze/Metadaten
- Zuweisen zu Glossary Assets

The screenshot displays the IBM InfoSphere Information Governance Catalog (IGC) interface. The top navigation bar includes 'Search', 'Glossary', 'Information Assets', 'Labels', 'Queries', and 'Collections'. The left sidebar shows a tree view of 'Implemented Data Resources' with 'ID_PARTNER' selected. The main panel shows 'DATABASE COLUMN DETAILS' for '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 – Funktionen

Queries

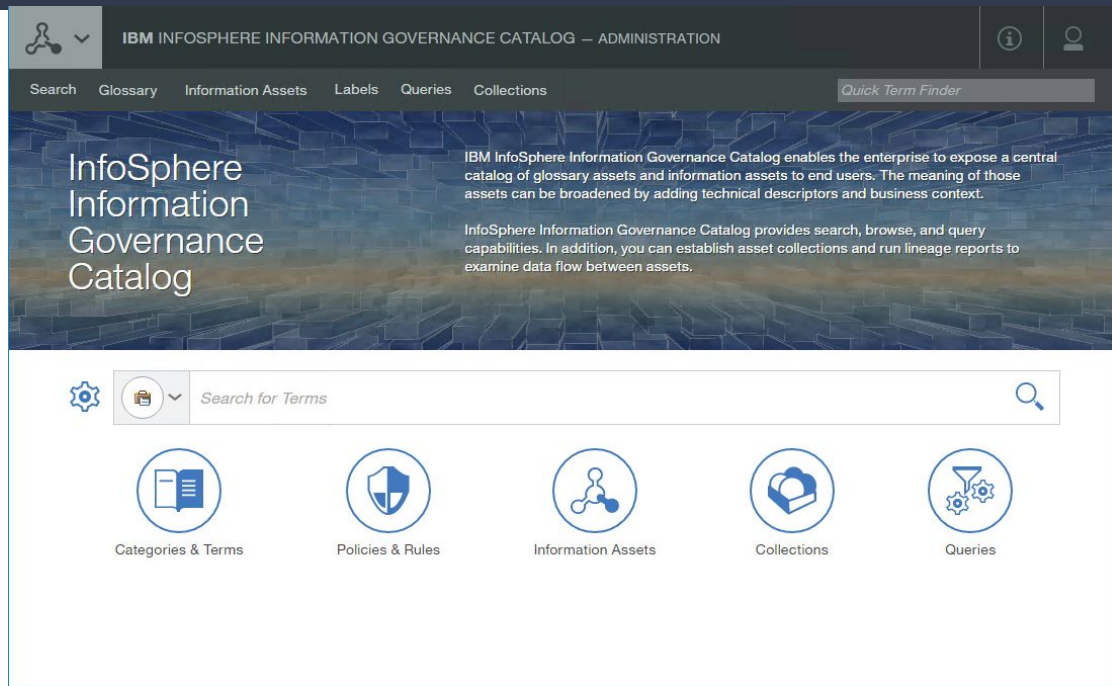
- Erstellen eigener Queries
- Für Information Assets
- Für Glossary Assets
- Ergebnis: Tabelle mit Infos

The screenshot displays the IBM InfoSphere Information Governance Catalog (IGC) interface. The top navigation bar includes links for Search, Glossary, Information Assets, Labels, Queries, and Collections. The main content area is titled 'Untitled Query' and features a 'Run' button, a 'Save' button, and a 'Clear' button. On the left, the 'ASSET TYPE' is set to 'Term', and the 'AVAILABLE PROPERTIES' list includes Name, Short Description, Long Description, and Parent Category. The 'DISPLAY' tab is active, showing a criteria table with three rows. The first two rows are grouped under a yellow header 'criteria must be met.' and the third row is under a blue header 'criteria must be met.'.

Criteria ID	Property	Operator	Value
1.1	NAME	Contains	PARTNER
1.2	NAME	Contains	MEMBER
2	PARENT CATEGORY - NAME	Contains	PARTNERWORLD

IBM InfoSphere IGC – Für wen geeignet?

- Business Analysten
- Experten der Fachbereiche
- Unternehmen, die
 - ein gemeinsames Unternehmens-Vokabular sowie Governance-Praktiken verwalten wollen
 - das Potentials integrierter Metadaten nutzen wollen
 - den Bedarfs an technischen Schulungen reduzieren wollen



Lumada Data Catalog (Waterline Data Catalog)

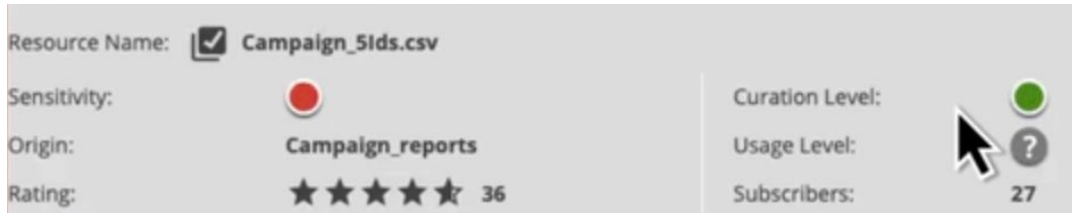
- Tool zur Verwaltung von Daten aus diversen Quellen
- nutzt Maschinelles Lernen, um Dateninventar aufzubauen
- patentierte Fingerabdruck-Technologie für Daten
- Verwaltung von Data Lakes
- Ziel: Große Datenmengen automatisch analysieren
- Preis und Demo sind nur auf Anfrage verfügbar

The screenshot displays the Lumada Data Catalog interface. The top navigation bar includes links for Home, Catalog (highlighted), Data Studio, and Edge, along with a user profile for Thomas Clark, Data Scientist. The main content area is titled 'Data Catalog' and features three summary cards: Assets (378), Events (234), and Insights (157). Below these is a search bar showing '2 of 123 items' and a 'Sort by' dropdown set to 'Newest first'. A table lists various data assets with columns for Name, Location, Entity Type, Tags, and Last Updated. The table includes items like 'Sales forecast model', 'Storage Flow HeatMap', 'Storage Entrance Video', 'Number of Visitors', and 'Sales Data'. A sidebar on the left allows for refining the search by Entity Type (Assets, Events, Insights, Visualization, Dataset, Video, Image, Structured) and Databub (Raw Zone, Curated Zone, Published Zone, Edge, Cloud). It also includes a Tags section with Customer and CRM filters.

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

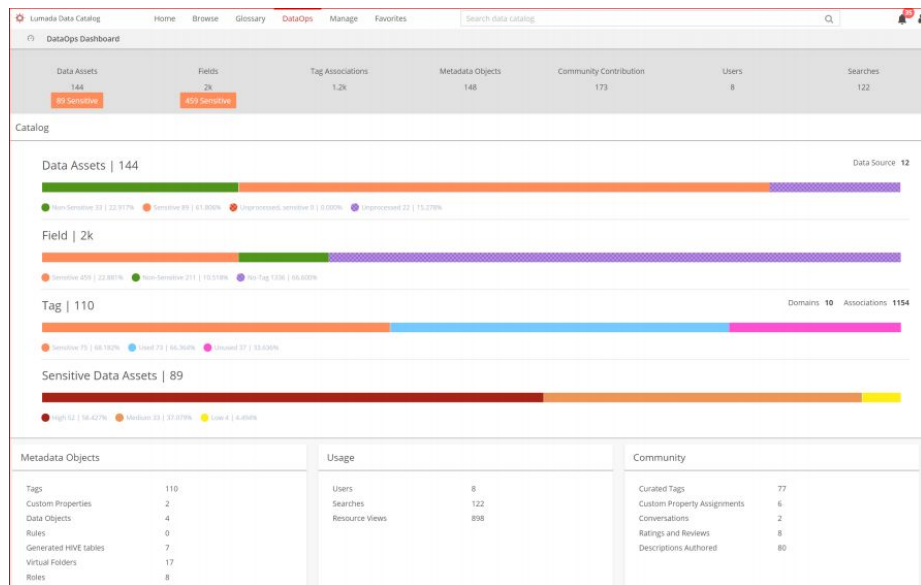
Lumada Data Catalog – Funktionen

- Datenerkennung mit Hilfe metadatenbasierter Suche
- Verwaltung von sensiblen Daten
- patentiertes Data Fingerprinting
 - Datenverarbeitung basierend auf maschinellem Lernen
 - automatisches Erkennen von vertraulichen Daten
 - ermöglicht eine Google-ähnliche Suche mit firmeninternen Bezeichnungen
- Visualisierung von Daten, -herkunft und -zusammenhänge
- Erkennen von redundanten Daten
- Kommentar-, Abo-, Bewertungsfunktion
- Aufzeichnen von Nutzerdaten



Lumada Data Catalog – Für wen geeignet?

- Unternehmen mit großen Datenmengen
- Datenanalysten
- schnelle und effiziente Anpassung an Datenschutzbestimmungen
- Verwaltung von sensiblen Daten



Vergleich

Kategorie	IBM InfoSphere IGC	Lumada Data Catalogue
Gemeinsamkeiten	<ul style="list-style-type: none">• Data Catalog Tools• individuelles Preisangebot• Nutzung verschiedener Datenquellen	
Oberfläche	Web Tool	keine Angabe (eventuell Desktop-Anwendung)
Schwerpunkt	Datenverwaltung	Datenanalyse
AI	Nein	AI-basierte Datenanalyse und Strukturierung
Produkt-Portfolio	IBM InfoSphere Family	Lumada Data Services
“Social-Media” Tools	Nein	Ja (Kommentare, Abo, Bewertung)

Quellen

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

VIELEN DANK FÜR EURE AUFMERKSAMKEIT

