

# IBM Watson Knowledge Catalog

*An intelligent, integrated enterprise data catalog and data governance platform*



IBM Cloud

# Watson Knowledge Catalog: Search and Explore

The screenshot displays the Watson Knowledge Catalog interface, specifically the 'Enterprise' catalog. The top navigation bar includes 'Catalogs / Enterprise' on the left and 'Add to Catalog' with a dropdown and a refresh icon on the right.

The main content area features a search bar with the placeholder 'What assets are you looking for?' Below the search bar are three tabs: 'Watson Recommends' (selected), 'Highly Rated', and 'Recently Added'. To the right of these tabs is an 'Expand' button.

A sidebar on the left contains a 'Filter' section with two expandable categories: 'Asset types' and 'Tags'. Under 'Asset types', there are checkboxes for Data Asset (39), Connection (11), Notebook (2), and Model (1). Under 'Tags', there are checkboxes for Insurance (17) and Great-Outdoors (13).

The central area shows a list of 'Available Assets' with 25 of 53 assets displayed. The list includes:

- 2017 J.D. Power U.S. Auto Clai... (Owner: OWN)
- Auto Insurance Claim (Owner: OWN)
- Auto Insurance Claims (Owner: OWN)
- Auto Insurance Customers (Owner: OWN)
- Auto Insurance Policies (Owner: OWN)

Below the asset list, there are three cards representing specific data assets:

- Insurance Driver**: Data Asset (1/0/0). Owner: Dirk deRoos, Added: Sep 12, 2018 9:32 AM. Insura... | Demo. Rating: 5 stars (yellow).
- Great Outdoors Retailer Sales Recent**: Data Asset (1/0/1). Owner: Ricardo Buglio, Added: Feb 20, 2018 2:44 PM. Great-... | Sales | Retail. Rating: 5 stars (yellow).
- Princeton Video Customer Sales**: Data Asset (1/0/1). Owner: Ricardo Buglio, Added: Oct 31, 2017 9:22 AM. Prince... | Sales. Rating: 5 stars (yellow).

At the bottom of the page, there is a 'Filter' button and a 'Available Assets' link.

# Watson Knowledge Catalog: Glossary Terms

Business Glossary / Sensitive Information

Overview Related content

## Sensitive Information

Sensitive information that is protected against unwarranted disclosure.

Term details	Description
Creator: Ricardo Buglio Date created: Oct 31, 2017 Last editor: Ricardo Buglio Last modified: Jan 10, 2019	<p><b>Description</b></p> <p>Sensitive information is data that must be protected from unauthorized access to safeguard the individual or organization. There are different types of sensitive information: Personal identifiable information (PII) is data that can be traced back to an individual and that, if disclosed, could compromise the privacy of that person. Such information includes biometric data, medical information, personally identifiable (PII) and unique identifiers such as passport or Social Security numbers. Threats include not only theft but also disclosure of personal information that the individual would prefer remained private, such as encrypted both in transit and at rest. Business information - Sensitive business information includes risk to the company in question if discovered by a competitor or the general public. Such information includes acquisition plans, financial data and supplier and customer information, among other possibilities. The amount of data generated by businesses, methods of protecting corporate information from unauthorized access, and the potential consequences of such access are becoming increasingly complex.</p>

Business Glossary

What are you looking for? Sort by: Term name

# A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Confidential Information

Data that is classified as confidential

Confidential

Last modified: Jan 2, 2018

Personally Identifiable Information

Data that is classified as Personally Identifiable Information (PII).

PII

Last modified: Apr 23, 2018

Sensitive Information

Sensitive information that is protected against unwarranted disclosure.

Sensitive

Last modified: Jan 10, 2019

Business Glossary / Sensitive Information

Overview Related content

## Sensitive Information

Archive

Related content by type	Related content															
	<p>Assets (9)</p> <table border="1"><thead><tr><th>ASSET NAME</th><th>CATALOG NAME</th><th>LAST MODIFIED</th></tr></thead><tbody><tr><td>Auto Insurance Customers</td><td></td><td>Mar 30, 2018</td></tr><tr><td>CUSTOMERS</td><td>Enterprise</td><td>Nov 9, 2018</td></tr><tr><td>CUSTOMER DEMOGRAPHIC</td><td>Enterprise</td><td>Jan 10, 2019</td></tr><tr><td>Insurance Driver</td><td>Knowledge Catalog</td><td>Nov 16, 2018</td></tr></tbody></table> <p>View all</p>	ASSET NAME	CATALOG NAME	LAST MODIFIED	Auto Insurance Customers		Mar 30, 2018	CUSTOMERS	Enterprise	Nov 9, 2018	CUSTOMER DEMOGRAPHIC	Enterprise	Jan 10, 2019	Insurance Driver	Knowledge Catalog	Nov 16, 2018
ASSET NAME	CATALOG NAME	LAST MODIFIED														
Auto Insurance Customers		Mar 30, 2018														
CUSTOMERS	Enterprise	Nov 9, 2018														
CUSTOMER DEMOGRAPHIC	Enterprise	Jan 10, 2019														
Insurance Driver	Knowledge Catalog	Nov 16, 2018														

# Watson Knowledge Catalog: Data Rules and Policies

Policy Manager / Data Privacy

+ Add ▾

Category

## Data Privacy

Description

Enterprise data privacy policies pertaining to data classified as Sensitive, Classified or Personally Identifiable Information that have to be protected and comply with government and industry regulations.

Creator: Ricardo Buglio    Created: Jul 31, 2017    Contains: 3 published policies and 4 rules

What are you looking for?

NAME	TYPE	STATUS	CONTAINS	LAST MODIFIED
Sensitive Information	Policy	Published	2 Items	Jan 10, 2019
Anonymize Government Identity Information	Rule	Published	--	Jan 10, 2019
Anonymize Financial Account Information	Rule		Data Dashboard	
Confidential Information	Policy			
Personally Identifiable Information	Policy			

Data Dashboard

Policy enforcements over time

Dec 16, 2018 - Jan 16, 2019

POLICY ENFORCEMENTS

Days

Jan 9, 2019  
Enforcements: 63

Scroll Down

Action \* ⓘ

then anonymize data in columns containing

Government Identities ✖

Select how to anonymize data:

Redact BEFORE 452-821-1120 Replace data with Xs.

Substitute BEFORE 452-821-1120 Replace data with values that don't match the original format.

Mask BEFORE 452-821-1120 Replace data with similarly formatted values.

Visit the documentation to learn more about data anonymization.

Data assets containing personal or restricted data

Confidential Information ⓘ 2 View all

Personally Identifiable Infor... ⓘ 2 View all

Sensitive Information ⓘ 9 View all

Operational policies ⓘ 3 Data policies

Automatic enforcement ⓘ 230 enforcements in last 30 days 71.71% from last month

# Watson Knowledge Catalog: Review and Understand

Catalogs / Enterprise / CUSTOMER\_DEMOGRAPHIC

+ Add to Catalog



Overview Access Review Profile Lineage

Data Asset

## CUSTOMER\_DEMOGRAPHIC

Remove Download Add to Project

### Description

Customer churn data asset

Added: Jan 10, 2019 11:09 AM

Format: CSV

Size: 322 KB

### Tags

Customer | Churn

### Reviews

1 reviews

Schema: 28 Columns | 2066 Rows | 2 Columns anonymized [\(i\)](#)

Preview: 1000 rows | Last refresh: 1 day ago | Refresh

EMAIL_ADDRESS	PHONE_NUMBER	NATIONAL_ID	CREDITCARD_NUMBER	CREDITCARD_TYPE	CREDITCARD_EXP
Type: String	Type: String	Type: String	Type: String	Type: String	Type: String
Email Address	US Phone Nu...	US Social Se...	Credit Card Number	Organization Na...	Date
wfronsekb1@source...	434-553-8337	231-24-6500	7df7ea4aac5e3493ea3127e	VISA	5/23
dcoyekj@pcworld.co...	760-277-6466	649-96-7557	db7d354d8a0d2042c969dd	Diners Club	3/21

Catalogs / Enterprise / CUSTOMER\_DEMOGRAPHIC

+ Add to Catalog



Overview Access Review Profile Lineage

Current profile

164 classifiers

Last profile

10 Jan 2019 - 11:43 am

[View Log](#)

Columns

Rows

2,066

Delete

Update Profile

ID

Type: Smallint

CUSTOMER

Type: Varchar

NAME

Type: Varchar

COUNTRY

Type: Varchar

FREQUENCY

FREQUENCY

FREQUENCY

FREQUENCY

FREQUENCY

FREQUENCY

FREQUENCY

1927-1954

OH15577

1491-1518

EZ71780

371-397

CB44654

1899-1925

BP26492

3222-3248

GE50763

529-555

QT25383

1411-1437

AU96286

3873-3898

JW9183

503-528

FH83537

100%

0%

0%

100%

0%

0%

0%

100%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

0%

# Watson Knowledge Catalog: Data Preparation & Data Science tools

My Projects / Transform Project / Great Outdoor Customer Order... / Data Refinery

+ Operation Code an operation to cleanse and shape your data

Data Profile Visualizations

**Input**

Columns COUNTRY

Chart types Pie Of Counts

Brunel syntax

```
bar y(#count) stack polar color(COUNTRY) sort(#count)
label(COUNTRY, '#, #count, "%') tooltip(#all) percent(#count)
legends none
```

Learn about our charts

A pie chart titled "Great Outdoor Customer Order" showing the distribution of customer orders by country. The data is as follows:

Country	Percentage
United States	17%
Germany	9%
China	8%
Brazil	7%
Japan	6%
Mexico	6%
United Kingdom	6%
Canada	4%
France	4%
Italy	4%
Singapore	4%
Australia	3%
Belgium	3%
Austria	2%
Netherlands	2%
Spain	2%

Details Help

DATA FLOW

Great Outdoor Customer Order  
Enter a description of the data flow

Steps: 0

Project Transform Project

DATA FLOW OUTPUT

Great Outdoor Cus

IBM Watson Studio Projects Tools Catalog Community Services Manage Support Docs

Location\* Transform Project, Data assets More

In [25]:

```
from pyspark.sql import functions as F
hourplot = collisions_final_df[['Bus','Truck','Taxi','Other','Hour','Auto']].groupBy('Hour')\
    .agg(F.sum("Bus").alias("Bus"), F.sum("Truck").alias("Truck"), F.sum("Taxi").alias("Taxi"),\ 
        F.sum("Other").alias("Other"), F.sum("Auto").alias("Auto")).toPandas()

hourplot[['Bus', 'Truck', 'Taxi', 'Auto']].plot(stacked=True, kind='bar', figsize=(12,8), alpha=1)
#SUM(vehicles)
plt.xlabel('Hour', size=17)
plt.ylabel('Vehicles', size=17)
plt.legend(loc='best', prop={'size':20}, framealpha=0)
plt.title('Collisions on Road per Hour', size=25)
plt.show()
```

Collisions on Road per Hour

A stacked bar chart titled "Collisions on Road per Hour" showing the number of collisions by vehicle type (Bus, Truck, Taxi, Auto) across different hours of the day. The Y-axis represents the number of vehicles (0 to 50,000) and the X-axis represents the hour (0 to 23). The chart shows two main peaks: one in the morning (around 6 AM) and one in the afternoon (around 5 PM). Buses, trucks, and taxis are involved in most collisions, while cars (Auto) are involved in fewer.

This plot shows collisions spread across a day, with peaks during the morning and afternoon rush hours. You can see that significantly more collisions occurred during the afternoon rush hour than during the morning rush hour. Also, the most collisions involve cars by far, while buses, taxis, and trucks are involved in accidents a lot less frequently.

# Business Problems Solved by WKC



## Problem #1: Inability to find the right data at the right time for reporting to the business and for creating models

**Persona:** Data Scientist or Business Analyst

**Solution:** Knowledge Catalog indexes and classifies data assets which can be quickly found by Intelligent search algorithms that can guide your data scientists and BAs to find the best data for their purpose. Once found they can move/share the data into their AI / Analytics projects.

**Value:** Improve the findability and use of data assets for business analysis or model development

## Problem #2: Inability to easily inventory data assets, document it and apply governance rules to protect data

**Persona:** Data Steward

**Solution:** Knowledge Catalog provides an easy way to build and index all assets across your business functions & then enforce access and anonymization using policies & rules.

**Value:** Intelligent data curation helps curate data faster and governed and protected data is made available to the business users to trust and use it with confidence.

## Problem #3: Information locked away in department silos

**Persona:** Data Scientist or Business Analyst

**Solution:** Knowledge Catalog makes it easy for even unstructured knowledge to be shared and accessible across the enterprise. ‘Auto Profiling’ and ‘data connections’ to 60+ sources makes it easy to catalog and organize. Cataloging of data creates a ‘shop for data’ online shopping experience experience.

**Value:** Find and use data through self-service access to the data

## Problem #4: Inability to deliver data efficiently and reliably to the business

**Persona:** Data Engineer

**Solution:** Knowledge Catalog provides self-service access to data. The business users no longer need to wait on IT to get the data they need and the Data Engineers can focus on the business critical tasks essential for business growth. CDO organization is directly in charge of the organization’s data mine.

**Value:** Quickly deliver data as and when new data sources are added to the data pipeline.

## Problem #5: Business and Software teams not in sync in a project

**Persona:** Application Developer / Business Analyst

**Solution:** IBM Knowledge Catalog provides an intelligent business terms and tags that connects business with software / IT orgs so, that all are on the same page.

**Value:** Commonality between business and software teams.

# Key Benefits



## Augmented Data Management

Automate the collection, curation and understanding of your data. Detect business entities, quality, bias and validity of your data to build an automatic level trust in your data.



## Data Prep & AI-powered Findability

Utilize automatic suggestions on how data can be best prepared for use. Intelligent suggestions to guide users to the best data for faster, smarter analytics and AI.



## Policy Activation for Data Governance

Autonomous enforcement of data and AI governance policies, providing automatic decisions to mask and protect data for a particular purpose.



## AI Governance

Support your adoption and roll out of trusted AI. Ensure that your models remain compliant for quality and bias.

***Automate the mundane***

IBM Cloud / March, 2019/ © 2019 IBM Corporation

***Use the data to get the data***

***Activate your governance program***

***Govern AI across your enterprise***