Conversational GUI for Semantic Automation Layer

Daniel Karl I. Weidele*, Gaetano Rossiello, Gregory Bramble, Abel N. Valente, Sugato Bagchi, Faisal Chowdhury, Mauro Martino, Nandana Mihindukulasooriya, Haritha Ananthakrishnan, Hendrik Strobelt, Hima Patel, Alfio Gliozzo, Owen Cornec, Ankush Gupta, Sameep Mehta, Manish Kesarwani, Horst Samulowitz, Oktie Hassanzadeh, Arvind Agarwal and Lisa Amini

IBM Research

Abstract

We present a Conversational Graphical User Interface for Semantic Automation Layer surfacing in *IBM watsonx.data Tech Preview*. Drawing from IBM pre-trained large language models (watsonx.ai), the layer semantically enriches data tables contained in a large, meta-structured data lakehouse (watsonx.data). We propose a graphical user interface backed by another fine-tuned watsonx.ai language model to provide business analysts with conversational access and control over the lakehouse using natural language.

Keywords

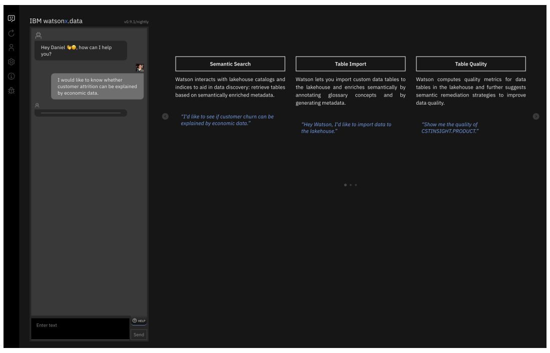
Conversational Graphical User Interface, Semantic Automation Layer, IBM, watsonx, Data Lakehouse, Semantic Data Science

1. Semantic Automation Layer

A semantic layer consists of business terms and glossary items that are meaningful in two main ways. Firstly, they make up the language of a particular business and secondly, they enhance existing data fields and labels that are cryptic and hard to understand. We build on the idea of a semantic layer as it is crucial for modern data management and data science, however, most of the semantic layers proposed in industry require constant human effort to ensure that the term mappings are correct for all of the data of the business. Coupled with the constantly changing nature of business data, this tedious manual process is both taxing to employees capacities and costly to businesses.

We therefore introduce a *Semantic Automation Layer* made up of four key components: (1) Semantic Search (keyword-based and search for joinables) using table information to build embeddings to more easily find and expand on enterprise data, (2) Semantic Enrichment that uses generative and discriminative artificial intelligence approaches to automatically caption, describe, tag and map tables and columns, (3) Glossary Utilities featuring semantically assisted techniques to merge and consolidate glossaries, to capture human feedback more natural and playful compared to the current industry practice, and (4) Data Quality Toolkit to provide data profiling and quality estimation metrics to assess the quality of ingested data in a systematic and objective manner.

ISWC 2023: International Semantic Web Confernce, November 06–10, 2023, Athens, Greece
*Corresponding author.
daniel.karl@ibm.com (D. K. I. Weidele)
0000-0001-5253-0511 (D. K. I. Weidele)


IBM watsonx.data

Hey Daniel 👋😊, how can I help you?

I would like to know whether customer attrition can be explained by economic data.

Semantic Search
Watson interacts with lakehouse catalogs and indices to aid in data discovery: retrieve tables based on semantically enriched metadata.
"I'd like to see if customer churn can be explained by economic data."

Table Import
Watson lets you import custom data tables to the lakehouse and enriches semantically by annotating glossary concepts and by generating metadata.
"Hey Watson, I'd like to import data to the lakehouse."

Table Quality
Watson computes quality metrics for data tables in the lakehouse and further suggests semantic remediation strategies to improve data quality.
"Show me the quality of CSTINSIGHT.PRODUCT."

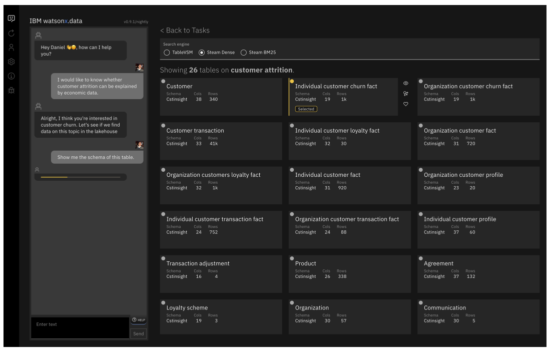
Figure 1: Main UI view comprising of chat component (left) and conversation starters (right).

2. Conversational GUI

We developed a Svelte web application to let users interact directly with the Semantic Automation Layer in a graphical way. The interface also comprises of a conversational agent in which a fine-tuned large language model can receive instructions from users (see figure 1).

To demonstrate business value of the overall system, let us assume the following user scenario: a business analyst (BA) would like to access and curate data from the lakehouse for further processing in their business intelligence application. BA specifically would like to investigate whether banking customer churn can be explained by economic data. The user phrases their instruction, upon which the conversational agent recognizes this as a search, retrieves data tables from the lakehouse and presents the results ranked by semantic relevance and with clean, AI generated table captions (see figure 2).

The conversational agent observes the UI context, so


IBM watsonx.data v0.9.1/nightly

Hey Daniel 👋😊, how can I help you?

I would like to know whether customer attrition can be explained by economic data.

Alright, I think you're interested in customer churn. Let's see if we find data on this topic in the lakehouse

Show me the schema of this table.

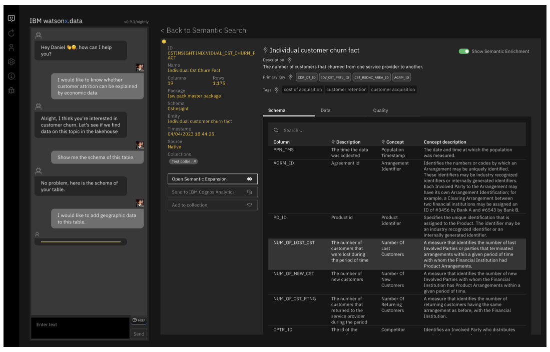
< Back to Tasks

Search engine: TableVSM [✓ Steam Dense] Steam BM25

Showing **26** tables on **customer attrition**.

Table	Schema	Cols	Rows
Customer	Cstinsight	38	340
Individual customer churn fact (Selected)	Cstinsight	19	1k
Organization customer churn fact	Cstinsight	19	1k
Customer transaction	Cstinsight	33	41k
Individual customer loyalty fact	Cstinsight	32	30
Organization customer fact	Cstinsight	31	720
Organization customers loyalty fact	Cstinsight	32	1k
Individual customer fact	Cstinsight	31	920
Organization customer profile	Cstinsight	23	20
Individual customer transaction fact	Cstinsight	24	752
Organization customer transaction fact	Cstinsight	24	88
Individual customer profile	Cstinsight	37	60
Transaction adjustment	Cstinsight	16	4
Product	Cstinsight	26	338
Agreement	Cstinsight	37	132
Loyalty scheme	Cstinsight	19	3
Organization	Cstinsight	30	57
Communication	Cstinsight	30	5

Figure 2: Semantic Search showing AI-captioned tables ranked by semantic relevance.


IBM watsonx.data v0.9.1/nightly

Hey Daniel 👋😊, how can I help you?

I would like to know whether customer attrition can be explained by economic data.

Alright, I think you're interested in customer churn. Let's see if we find data on this topic in the lakehouse

Show me the schema of this table.

No problem, here is the schema of your table.

I would like to add geographic data to this table.

< Back to Semantic Search

ID: CSTINSIGHT.INDIVIDUAL_CST_CHURN_FACT
Name: Individual Cst Churn Fact
Columns: 19 Rows: 1,175
Package: Isw pack master package
Schema: Cstinsight
Entity: Individual customer churn fact
Timestamp: 04/04/2023 18:44:25
Source: Native
Collections: Test collie

Individual customer churn fact — Show Semantic Enrichment

Description: The number of customers that churned from one service provider to another.

Primary Key: CDR_DT_ID, IDV_CST_PRFL_ID, CST_RSDNC_AREA_ID, AGRM_ID

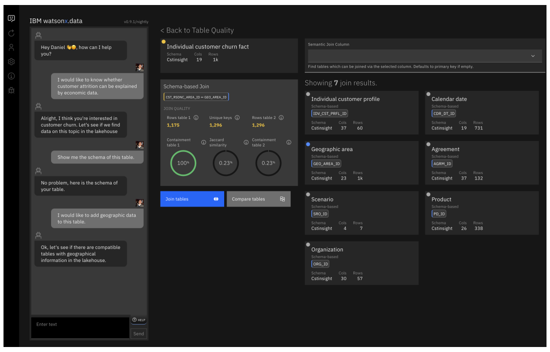
Tags: cost of acquisition, customer retention, customer acquisition

Column	Description	Concept	Concept description
PPN_TMS	The time the data was collected	Population Timestamp	The date and time at which the population was measured.
AGRM_ID	Agreement id	Arrangement Identifier	Identifies the numbers or codes by which an Arrangement may be uniquely identified. These identifiers may be industry recognized identifiers or internally generated identifiers. Each Involved Party to the Arrangement may have its own Arrangement Identification; for example, a Clearing Arrangement between two financial institutions may be assigned an ID of #3456 by Bank A and #6543 by Bank B.
PD_ID	Product id	Product Identifier	Specifies the unique identification that is assigned to the Product. The identifier may be an industry recognized identifier or an internally generated identifier.
NUM_OF_LOST_CST	The number of customers that were lost during the period of time	Number Of Lost Customers	A measure that identifies the number of lost Involved Parties or parties that terminated arrangements within a given period of time with whom the Financial Institution had Product Arrangements.
NUM_OF_NEW_CST	The number of new customers	Number Of New Customers	A measure that identifies the number of new Involved Parties with whom the Financial Institution has Product Arrangements within a given period of time.
NUM_OF_CST_RTNG	The number of customers that returned to the service provider during the period	Number Of Returning Customers	A measure that identifies the number of returning customers having the same arrangement as before, with the Financial Institution.
CPTR_ID	The id of the	Competitor	Identifies an Involved Party who distributes

Figure 3: Semantically enriched table with auto-generated caption, description, tags and estimated primary key, as well as column descriptions and glossary mappings.

when BA requests to take a look into the schema of *this table*, agent knows which table is meant. BA also could have opened the table via a direct click, which would have auto-generated a language utterance to still keep chat history and UI state in sync. Figure 3 shows the requested schema of the *Individual customer churn fact* table including semantic table enrichments (each indicated with a light bulb icon) from top to bottom: auto-caption, generated description, estimated primary key, generated tags. On the column level, the Semantic Automation Layer further generates descriptions for otherwise cryptic column names, and maps columns to business glossary concepts.

Let us further assume BA is in possession of a local CSV file, which contains unemployment rates for different geographies. To later join this local CSV file with the customer churn table via zip codes, BA first requests to enrich the churn table to see if geographic information can be added to this table. The agent recognizes this intent as a search for joinable tables. BA selects the *Geographic area* table and explores join quality metrics, such as the identified join column and containment or Jaccard coefficients.


IBM watsonx.data v0.9.1/nightly

Hey Daniel 👋😊, how can I help you?

I would like to know whether customer attrition can be explained by economic data.

Alright, I think you're interested in customer churn. Let's see if we find data on this topic in the lakehouse

Show me the schema of this table.

No problem, here is the schema of your table.

I would like to add geographic data to this table.

Ok, let's see if there are compatible tables with geographical information in the lakehouse.

< Back to Table Quality

Individual customer churn fact — Schema: Cstinsight, Cols: 19, Rows: 1k

Schema-based Join: CST_RSDNC_AREA_ID ↔ GEO_AREA_ID

JOIN QUALITY

Rows table 1	Unique keys	Rows table 2
1,175	1,296	1,296

Containment table 1	Jaccard similarity	Containment table 2
100%	0.23%	0.23%

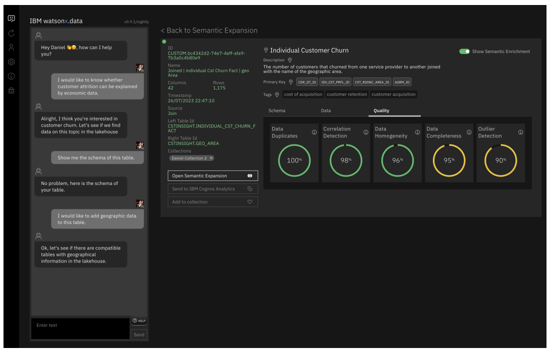
Semantic Join Column

Find tables which can be joined via the selected column. Defaults to primary key if empty.

Showing **7** join results.

Table	Join	Key	Schema	Cols	Rows
Individual customer profile	Schema-based	IDV_CST_PRFL_ID	Cstinsight	37	60
Calendar date	Schema-based	CDR_DT_ID	Cstinsight	19	731
Geographic area	Schema-based	GEO_AREA_ID	Cstinsight	23	1k
Agreement	Schema-based	AGRM_ID	Cstinsight	37	132
Scenario	Schema-based	SRO_ID	Cstinsight	4	7
Product	Schema-based	PD_ID	Cstinsight	26	338
Organization	Schema-based	ORG_ID	Cstinsight	30	57

Figure 4: Semantic Join Search with join quality metrics.


IBM watsonx.data v0.9.1/nightly

Hey Daniel 👋😊, how can I help you?

I would like to know whether customer attrition can be explained by economic data.

Alright, I think you're interested in customer churn. Let's see if we find data on this topic in the lakehouse

Show me the schema of this table.

No problem, here is the schema of your table.

I would like to add geographic data to this table.

Ok, let's see if there are compatible tables with geographical information in the lakehouse.

< Back to Semantic Expansion

ID: CUSTOM.bc4342d2-74e7-4eff-afa9-7b3a0c4b80e9
Name: Joined | individual Cst Churn Fact | geo Area
Columns: 42 Rows: 1,175
Timestamp: 26/07/2023 22:47:10
Source: Join
Left Table Id: CSTINSIGHT.INDIVIDUAL_CST_CHURN_FACT
Right Table Id: CSTINSIGHT.GEO_AREA
Collections: Daniel Collection 2

Individual Customer Churn — Show Semantic Enrichment

Description: The number of customers that churned from one service provider to another joined with the name of the geographic area.

Primary Key: CDR_DT_ID, IDV_CST_PRFL_ID, CST_RSDNC_AREA_ID, AGRM_ID

Tags: cost of acquisition, customer retention, customer acquisition

Data Duplicates	Correlation Detection	Data Homogeneity	Data Completeness	Outlier Detection
100%	98%	96%	95%	90%

Figure 5: Data quality of joined table

The resulting table (see figure 5) contains 42 columns including the 23 additional geographical area features. BA can further explore the data quality report for the joined table, comprising of metrics such as Data Duplicates, Data Completeness, Outlier Detection, and more.

In a last step, which is not represented graphically here, BA imports their unemployment data CSV into the lakehouse, before performing a final join with the churn-geo table via zip codes.

3. Conversational Agent

The conversational agent is a fine-tuned JSON-to-JSON transformer model which receives the current UI view ID, the context map containing user selections or other UI observations, as well as the user utterance as its JSON input string. The goal of the model is to perform three tasks (1) intent recognition, (2) entity extraction and (3) response generation in a single pass. The output of the transformer model therefore is a JSON string containing the view ID mapping to the user's intent, the required input parameters for the view, as well as a friendly response to be presented in the chat component.