

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 Strobel, Hima Patel, Alfio Gliozzo, Owen Corne, 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

ISWC 2023: International Semantic Web Conference, November 06–10, 2023, Athens, Greece


*Corresponding author.

✉ daniel.karl@ibm.com (D. K. I. Weidele)

ORCID 0000-0001-5253-0511 (D. K. I. Weidele)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

profiling and quality estimation metrics to assess the quality of ingested data in a systematic and objective manner.

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 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.

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.

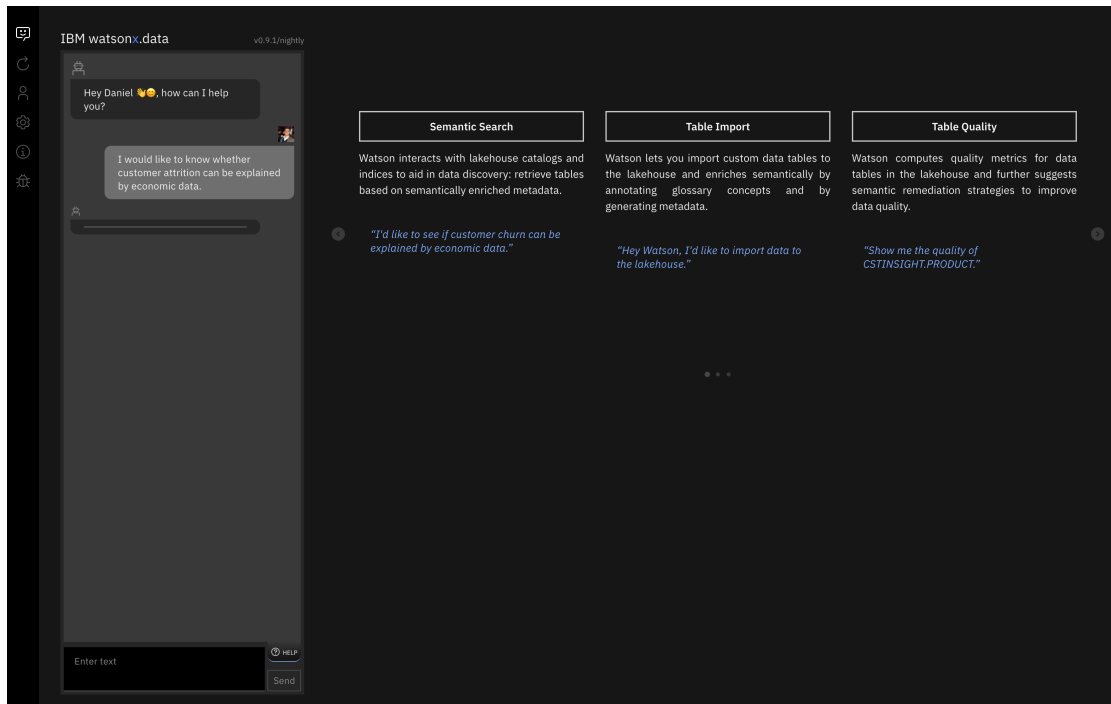


Figure 1: Main UI view comprising of chat component (left) and conversation starters (right).

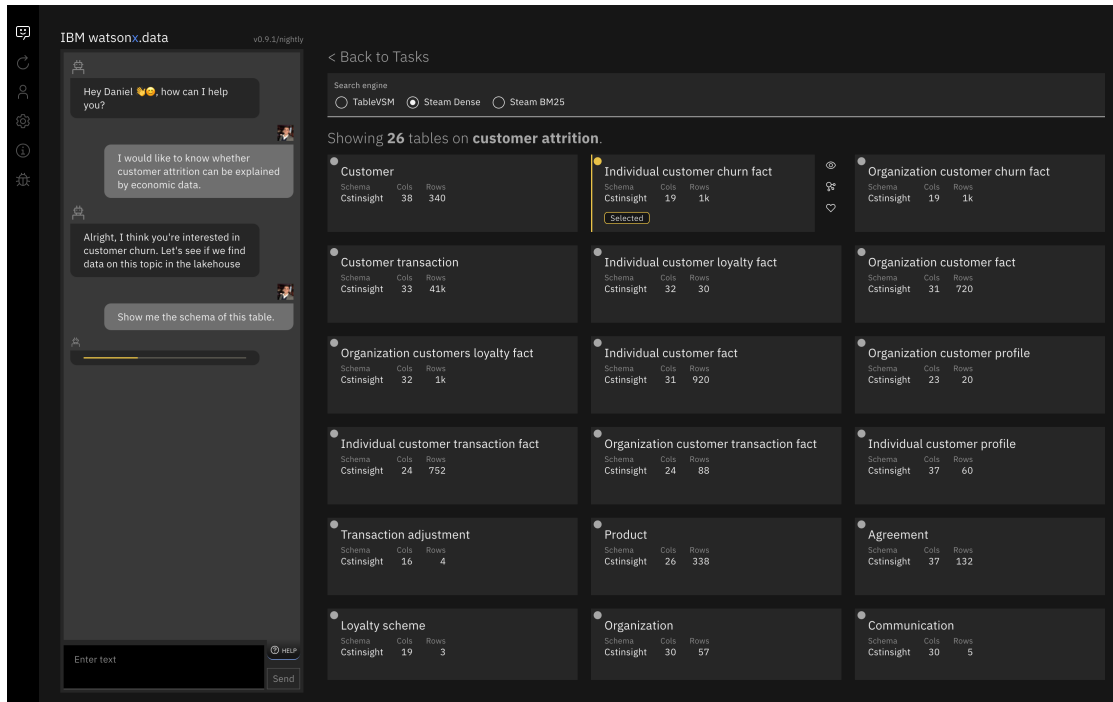


Figure 2: Semantic Search showing AI-captioned tables ranked by semantic relevance.

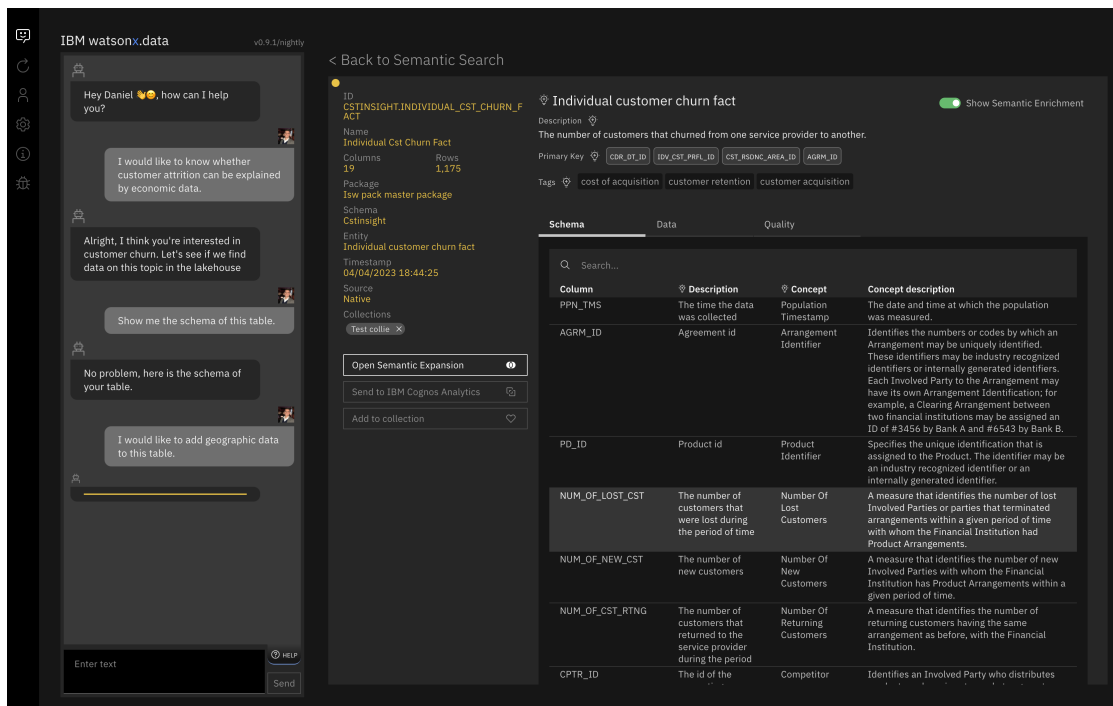


Figure 3: Semantically enriched table with auto-generated caption, description, tags and estimated primary key, as well as column descriptions and glossary mappings.

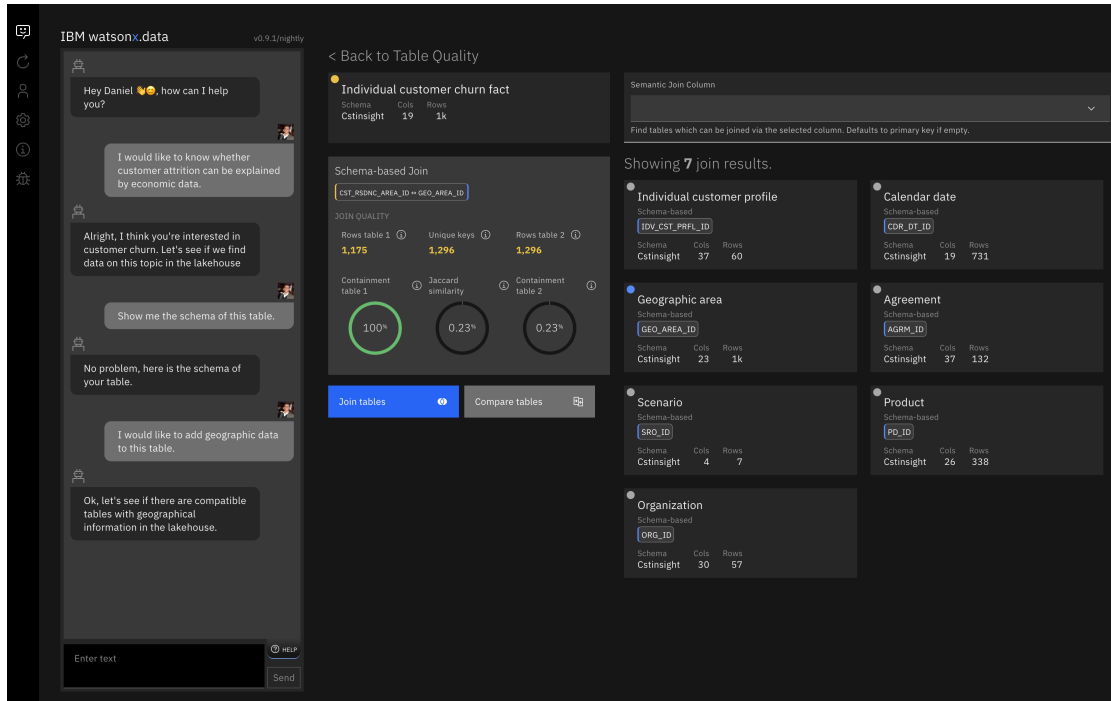


Figure 4: Semantic Join Search with join quality metrics.

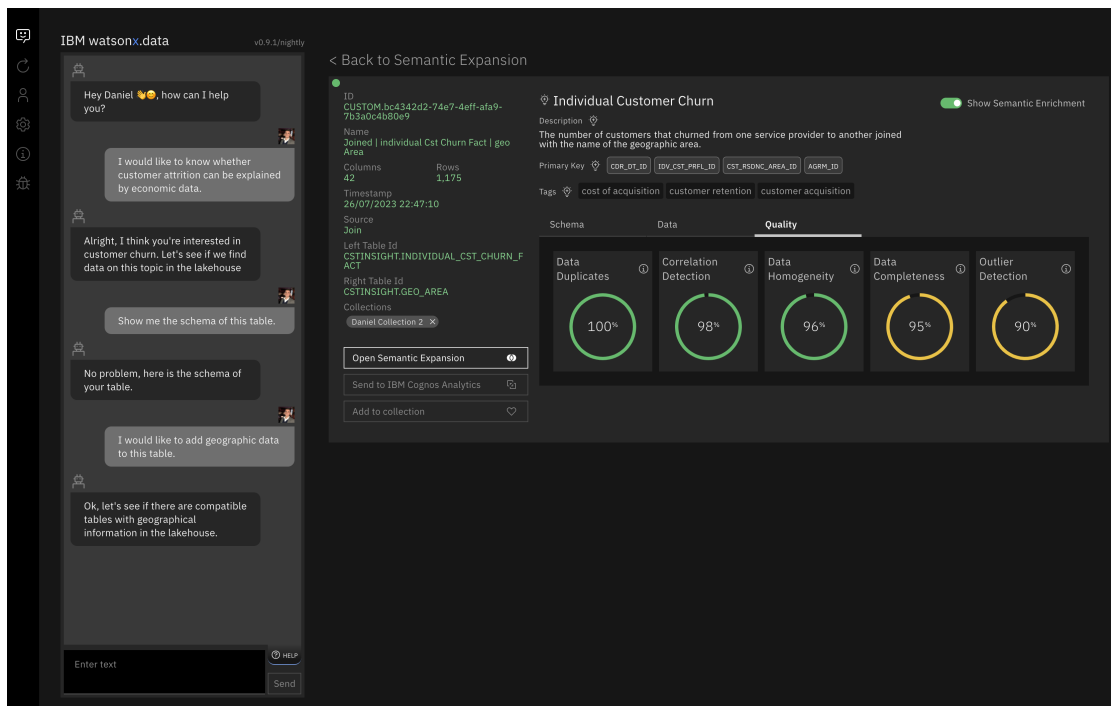


Figure 5: Data quality of joined table