SQL query generation based on natural language questions

Motivation

Data professionals need to possess extensive knowledge and experience in multiple domains, know specific definitions and understand business processes and requirements, to generate accurate SQL queries. This creates a significant barrier to entry for non-technical stakeholders who need to access data insights, with challenges like:

- Incomplete documentation
- Ambiguity of natural language
- Requirement to combine multiple data sources that were not designed to be evaluated in combination (no common identifier, …)

Main objective & tasks

The main objective is to support non-technical stakeholders in creating SQL queries for complex cross-domain datasets.

Main task

Develop a tool to translate natural language questions into SQL queries providing answers.

You can assume that the user asking the natural language is a business expert without any knowledge of SQL or databases. You may also assume that questions are phrased so that there is a unique correct way to answer them (i.e. no conversation or clarifications required).

Questions to tackle

- What data structures and formats work well for making use of the available documentation?
- Which architecture makes sense for such an application? (No frontend required)
- How can the quality of the results be evaluated?

Example for natural language questions

- *Which application had the largest relative revenue growth from calendar years 2022 to 2024?*
- *What are the three packages with the highest number of products sold across PL15 and PL18 in the last three years and how does the revenue distribute between the two PL across these three packages?*
- *What is the total revenue for customer X?*
- *For which products has the yearly revenue been decreasing year-on-year for the last 3 financial years, showing those revenues and the relative decrease?*

Bonus task: Clarify ambiguities

- While for the examples above there is a unique solution, one could also ask questions that require additional clarification. As a bonus task, a conversational approach clarifying potential ambiguities could be implemented. Again, the target users are business experts without knowledge of SQL or databases.

Examples for questions with ambiguities

- *How many pieces of product X have been sold in 2023? (→ calendar or fiscal year?)*
- *Which was the biggest customer of PL51 in FY23/24 by the total number of sold pieces? (→ how do we define customer?)*

Restrictions

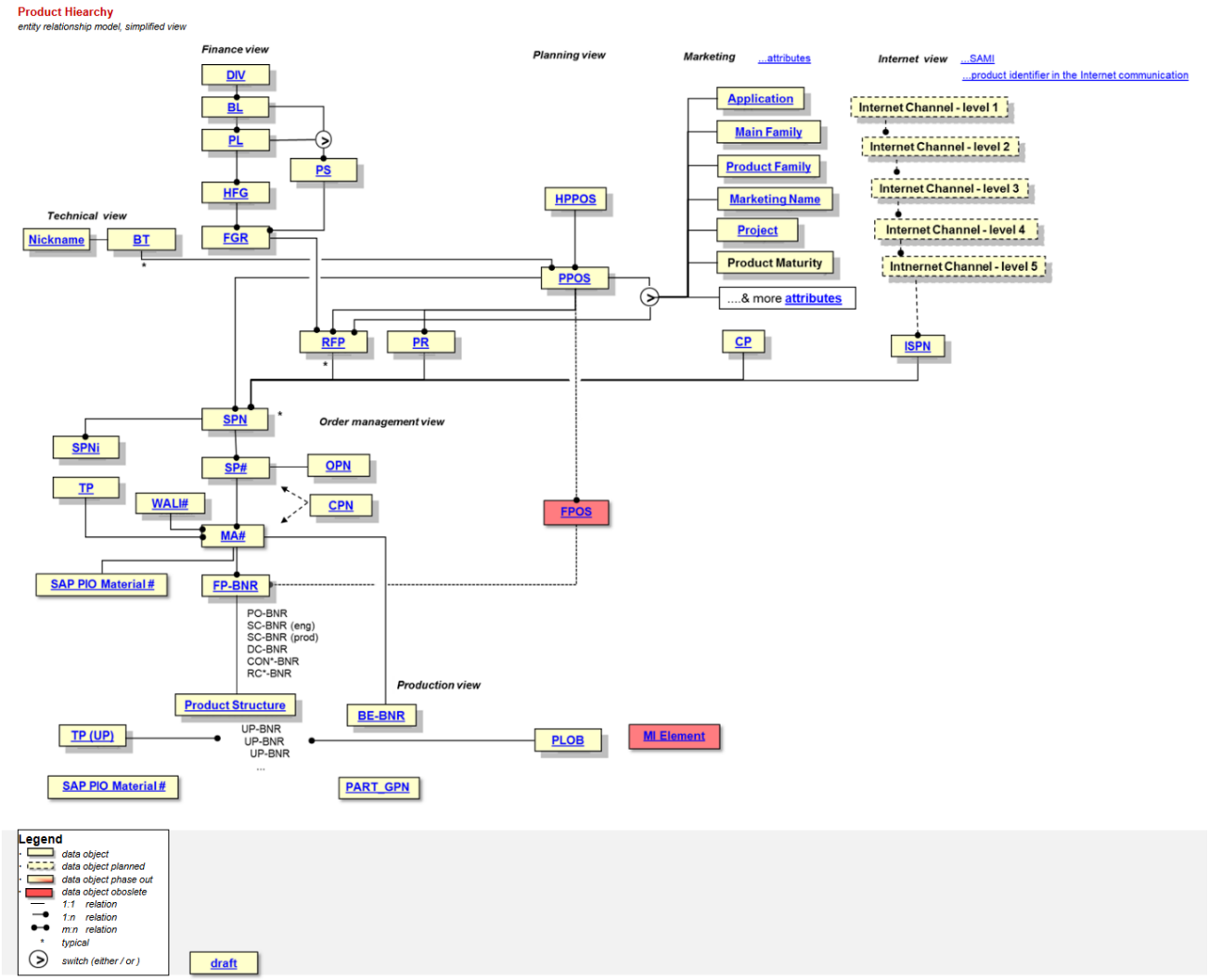
- Data must not be shared.
- Ensure to not share sensitive information within your prompts.
- The solution should work using one (or multiple) of those models:
    - DeepSeek 70b
    - GPT4o
    - GPT4o mini
    - Llama 3.1 70b
    - Llama 3.3 70b
    - Mixtral
    - O3mini
- The generated query should use syntax suitable for mySQL

Infineon supervisors

Anna Teiwes

Documentation

1 Product Hierarchy



2 Description of tables

table_name	table_description
bv_cl_fin_mkt_biz_pgs_hierarchy_cleaned	This table contains information about product hierarchies, such as the product line, the business line and many more
bv_cl_fin_mkt_biz_pgs_package_cleaned	This table contains general information about packaging, such as the manufacturer, the package material, the package technology, the package weight and more
bv_cl_fin_mkt_biz_pgs_sales_product_cleaned	This table contains information about the sold products, such as product category, sales product name, sales product number and more

fin_data_pss_extract	This table contains a lot of financial information, such as the financial year/quarter/month, the business line, the product line
bv_cl_fin_mkt_biz_sami_elec_params_pss_cleaned	This table contains technical test information about hardware products, such as the product category, the product name, the voltage class, the maximum tested voltage and more

3 Description of columns

See the attached Excel file (*column_descriptions.xlsx*)

4 Abbreviations

Abbreviation	Full Term	Comment
DIV	Division	
BL	Business Line	
PL	Product Line	
HFG	Hauptfabrikategruppe	German for Main Product Group.
FGR	Fabrikategruppe	German for "Product Group".
SPN	Sales Product Name	
SP#	Sales Product Number	
PPOS	Planposition	Product-related planning variable suitable for sales and marketing

5 Specific business knowledge

There are some specific terms and definitions that are important to consider in order to translate the questions correctly.

- **Fiscal year**
  - Unlike the calendar year, our fiscal year starts in October
  - Fiscal year 24/25 is defined as October 2024 – September 2025
  - To allow fiscal months to be date types, fiscal year 24/25 is often referred to as 2025 in the data, i.e. fiscal month 2025-03 is the third month of fiscal year 2024/2025, i.e. December 2024 (calendar year/month).
- **Organizational structure**
  - As shown in the product hierarchy diagram, there are multiple organizational levels within a business division.
  - The data only contains the division PSS.

- For every DIV, BL, PL, HFG and FGR, there are multiple ways for identification:
  - There is a name describing them
  - There is an abbreviation of the name
  - There is a number identifying them
- Example for one specific PL:
  - Medium Voltage Switches
  - MVS
  - 51
  - Depending on the data source, this specific PL could be identified in different ways; it may be necessary to join "*PL51*" (string) to *"51_MVS"* (string) or to *51* (integer) when combining multiple tables.
- In natural language (i.e. in the input questions you can expect, but ideally also in the output you create), the following conventions are common:
  - Divisions are revered to with their abbreviation, for example PSS
  - BLs are referred to by either full name or abbreviation, for example Power Switches or PS
  - PLs are referred to by their number with the prefix "PL", for example PL51
  - HFG and FGR are referred to by their full name

- **Customers**
  - While many big customers buy products directly with us, another part of our business is sales through distribution channels. This is very typical for the semiconductor industry, where many companies buy products through distributors, often also referred to as "disty" or "disti".
  - This leads to the fact that for one single sales transaction, there is more than one customer involved.
  - There are different ways to define customers:
    - Main customer
    - Pricing relevant customer
    - ..
  - Some customers buy through multiple channels, directly and via distribution
- Financial figures
  - There are multiple different measures for revenue, capturing different perspectives:
    - Estimates for revenue (rev_est) are entered for the future; for the past, they should correspond to the actual revenue.
    - For cost and margin calculations, a specific subset of the revenue (rev_mat) is relevant.
    - If we ask for revenue without any specific further information, we are interested in total sales to third parties (domestic and foreign customers) in a specific period in Euro including all order types.