SAP AI Query Generation Automation – Ground Truth System Proposal

1. Overview

This document outlines the proposed architecture and workflow for automating SQL query generation from SAP source systems using AI agents. The key goal is to eliminate manual intervention in generating join queries by leveraging semantic schema summaries and AI-based ground truth generation.

2. Objective

The system aims to automatically generate, validate, and store SAP SQL join queries using AI agents. Instead of relying on manual query writing or repeated schema inference, the process will generate a one-time 'Ground Truth' metadata structure that can be used for consistent, high-confidence query generation.

3. System Architecture

The architecture consists of six key components as shown in the diagram below:

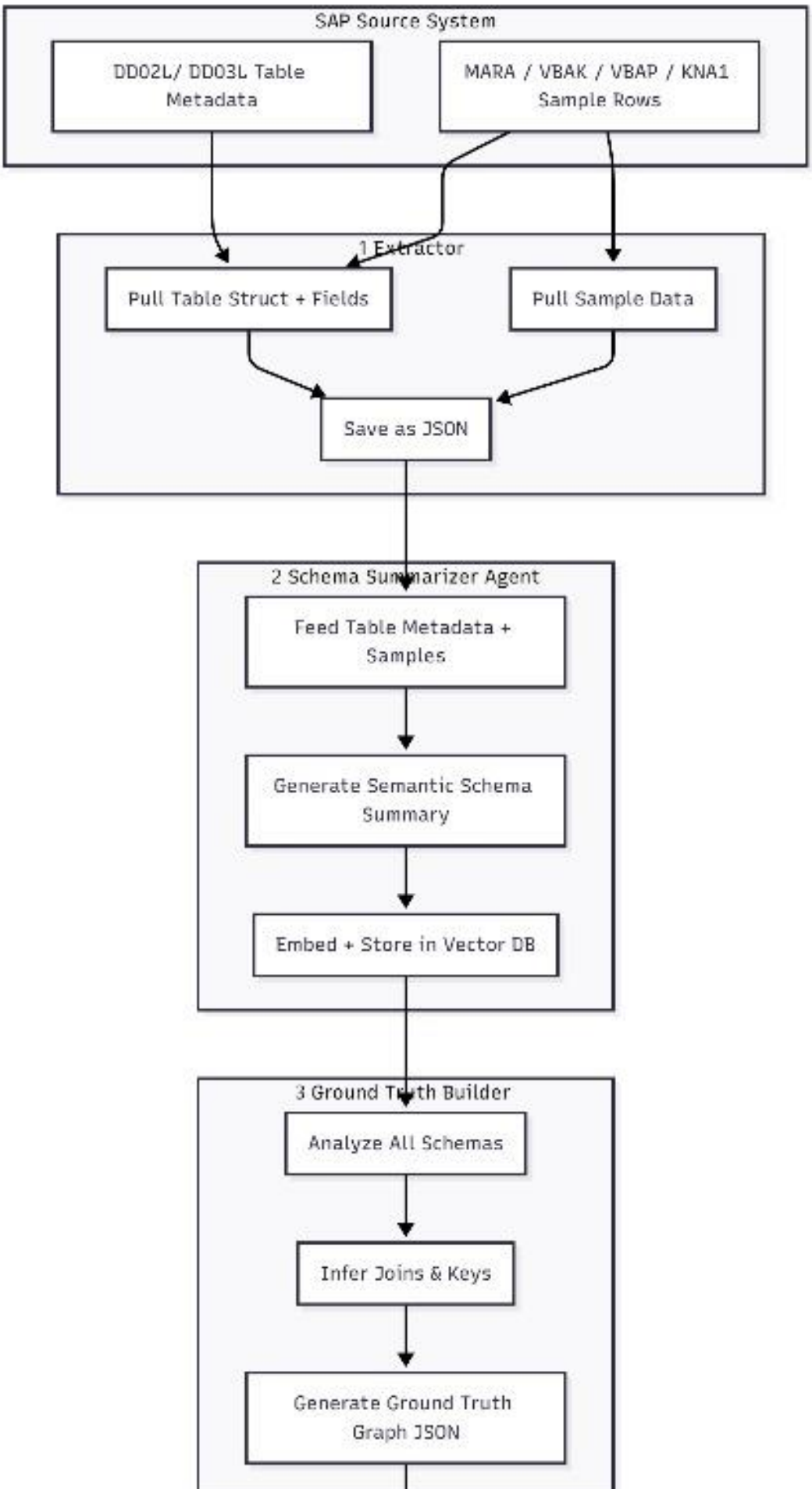
SAP Source System
DD02L/ DD03L Table Metadata
MARA / VBAK / VBAP / KNA1 Sample Rows
1 Extractor
Pull Table Struct + Fields
Pull Sample Data
Save as JSON
2 Schema Summarizer Agent
Feed Table Metadata + Samples
Generate Semantic Schema Summary
Embed + Store in Vector DB
3 Ground Truth Builder
Analyze All Schemas
Infer Joins & Keys
Generate Ground Truth Graph JSON

Figure 1: Automated SAP Query Generation System Architecture.

4. Workflow Breakdown

The workflow involves six AI-driven stages:

- 1. Extractor:

Pulls table structure and sample data from SAP source tables (DD02L/DD03L for metadata, MARA/VBAK/VBAP/KNA1 for data). Data is saved as JSON for downstream processing.

- 2. Schema Summarizer Agent:

Generates a semantic summary for each table schema and embeds it in a vector database for semantic retrieval.

- 3. Ground Truth Builder:

Analyzes schemas to infer joins and primary/foreign keys, producing a consolidated ground truth graph in JSON.

- 4. Query Generator Agent:

Takes user prompts/use cases and generates synthetic SQL joins using the ground truth graph.

- 5. Validator Agent:

Cross-checks generated joins against the ground truth graph to ensure correctness and assigns confidence scores.

- 6. Output Store:

Saves the final JSON query with its confidence score and associated schema links for future retrieval.

5. Example Ground Truth JSON

Below is an example representation of the generated ground truth metadata structure:


```
{
 "tables": {
  "MARA": {
   "key": ["MATNR"],
   "fields": ["MATNR","MTART","MATKL","MEINS","LAEDA"],
   "delta_by": "LAEDA"
  },
  "VBAK": {
   "key": ["VBELN"],
   "fields": ["VBELN","AUART","ERDAT","KUNNR","VKORG"],
```

```
    "delta_by": "ERDAT"
   },
   "VBAP": {
    "key": ["VBELN","POSNR"],
    "fields": ["VBELN","POSNR","MATNR","KWMENG","WERKS"],
    "delta_by": "ERDAT"
   },
   "KNA1": {
    "key": ["KUNNR"],
    "fields": ["KUNNR","LAND1","ORT01","NAME1","REGIO"],
    "delta_by": null
   }
  },
  "joins": [
   {"left": "VBAP.MATNR", "right": "MARA.MATNR", "type": "inner"},
   {"left": "VBAP.VBELN", "right": "VBAK.VBELN", "type": "inner"},
   {"left": "VBAK.KUNNR", "right": "KNA1.KUNNR", "type": "left"}
  ]
}
```


6. Advantages

- Reduces token and compute costs by reusing schema-level embeddings and ground truth metadata.
- Ensures consistency in query generation across sessions and datasets.
- Allows for easy manual corrections or schema updates without reprocessing all data.
- Improves validation through automated confidence scoring and cross-verification.

7. Approval Request

The proposed system design is ready for internal review and approval. Once approved, the next steps include:

- • Implementing Extractor and Schema Summarizer modules.
- • Defining schema-to-vector embeddings and join detection heuristics.
- • Deploying the Query Generator and Validator agents for internal testing.

Prepared by: Veloria Labs
Date: October 2025