Cross-Platform Data Migration with Schema Evolution and LLM-Integrated QA Automation

Sourav Ghosh (M22AIE241)

Indian Institute of Technology Jodhpur

Introduction

- Modern data systems demand both scalability and intelligence to support growing data volumes and complex analytical needs.
- Key challenges include handling schema evolution, real-time data processing, and achieving efficient cloud-native scalability.
- This work focuses on cross-platform data migration, automated schema evolution, and QA automation powered by LLM-enhanced intelligence.
- I propose a Hybrid Retrieval-Augmented Generation (RAG) framework capable of reasoning over relational data, performing numerical and quantitative analysis, and delivering precise, contextually grounded answers.

Background of the work

- Traditional vs. Modern Data Systems: Traditional pipelines rely on rigid schemas and batch-oriented workflows, whereas modern architectures embrace flexible schemas, continuous ingestion, and real-time analytics.
- Dynamic Schema Evolution: Modern platforms support automated schema detection, drift handling, and seamless evolution without breaking downstream consumers.
- QA Frameworks & Reliability: Robust quality-assurance mechanisms ensure schema validation, anomaly detection, and proactive error resolution across heterogeneous data sources.
- LLM Integration for Intelligence: Large Language Models enable dynamic query generation, contextual reasoning, and actionable insights—bridging human intent with complex relational datasets.

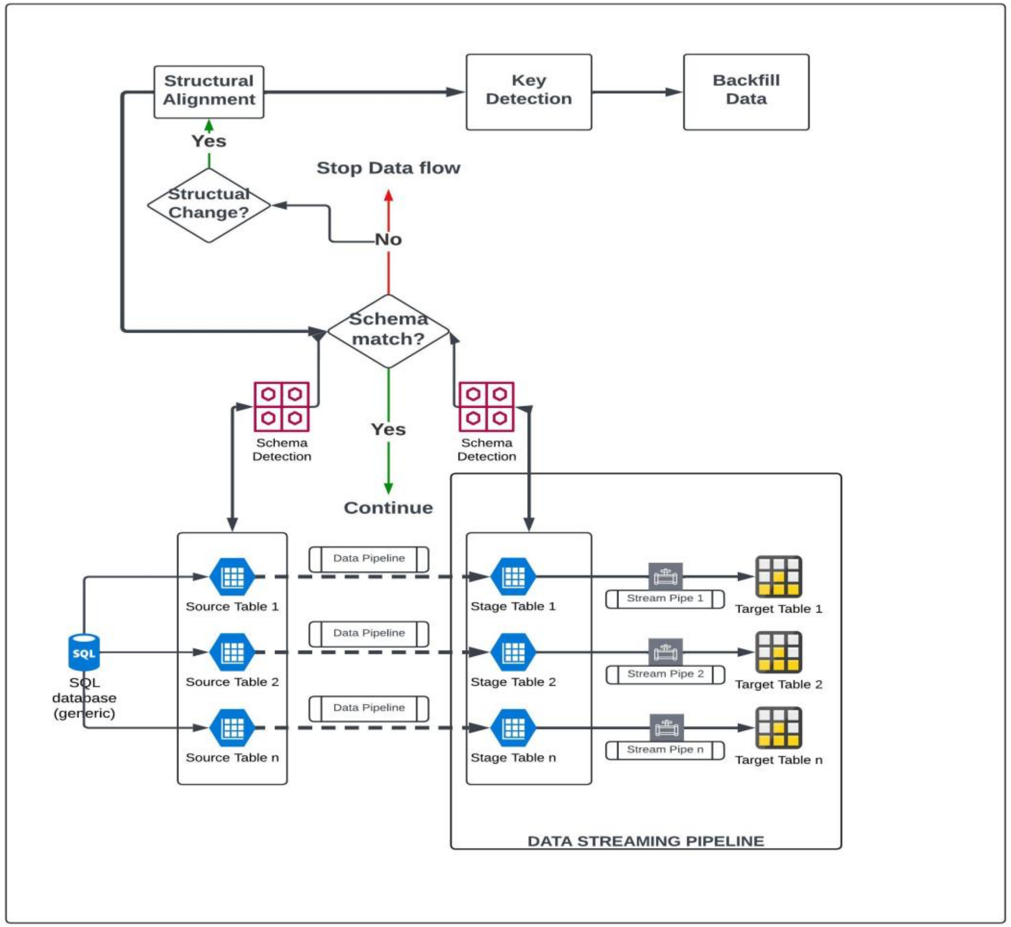
Problem Definition

- Cross-platform migration with automated schema mapping.
- Schema evolution tracking with version-aware transformations.
- LLM-powered QA automation for rule validation, and test-case generation.
- Hybrid RAG : LLM-based query generation & inference, enabling both quantitative analytics and natural-language insights directly from relational data.

Architecture

- Automated Schema Detection & Alignment:
  Intelligent identification of schema drift, column-level mismatches, and structural variations across cross-platform systems.
- Real-Time Data Migration with Historical Lineage Tracking:
  Continuous synchronization between heterogeneous platforms using CDC, versioned schema history, and audit trails.
- LLM-Enhanced QA Automation:
  AI-driven test-case generation, anomaly detection, rule-based validation, and natural-language QA for completeness and accuracy.
- Hybrid RAG over Relational + Unstructured Data:
  Enables LLM-powered SQL generation, query rewriting, cross-table reasoning, and inference over factual + contextual data.

Schema Detection and Evolution



1.Streaming pipeline will have the capability to trace insertion of new data or any update made to existing records.
Usage : Snowflake Streaming pipeline , Snowflake Task

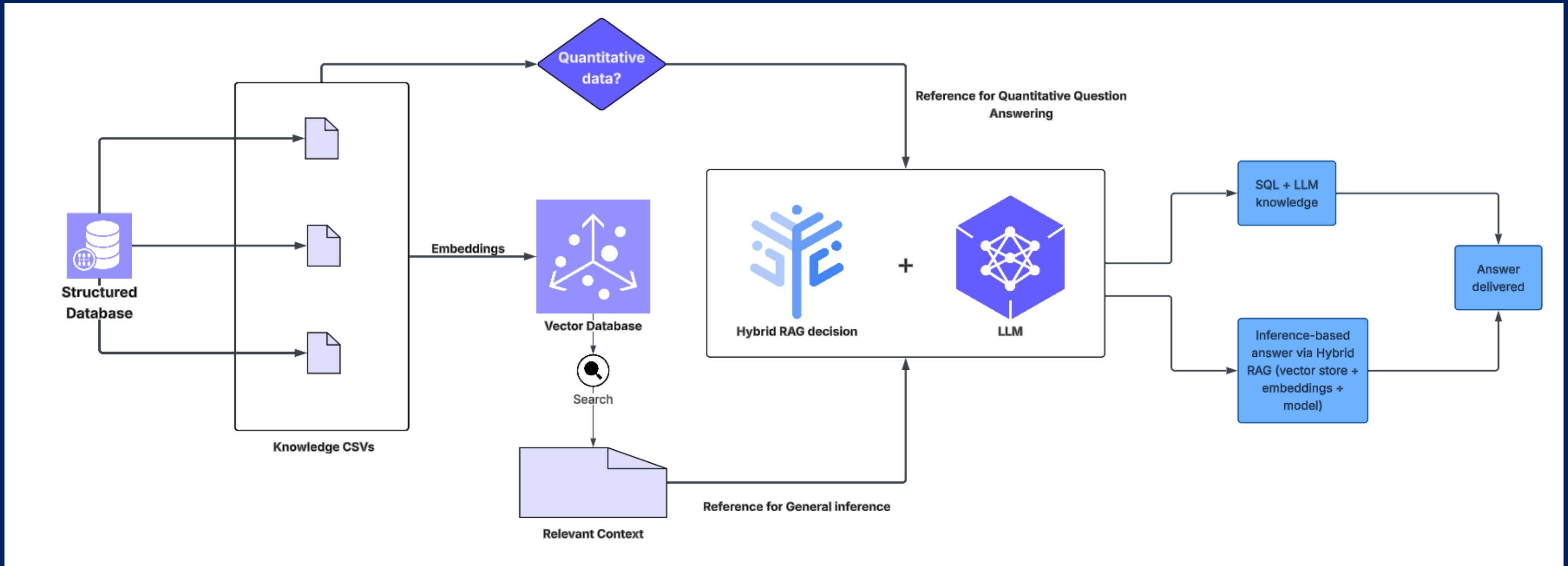
2.Snowflake platform stream can be used to periodically check for inserted records or updated records and merge to existing target tables.

3.Data PIpeline between Source and Stage : Fivetran or Matillion or Python scripting

SQL database (generic) — Any realtional Database system

Schema Detection — Kafka Confluent or Python Scripting

Hybrid RAG System



Hybrid RAG System

- User Query: User asks a natural-language business question.
- RAG Retrieval: Question is embedded → Chroma returns the most relevant context.
- Decision Engine: If the query needs numbers, filters, or aggregation → SQL mode. If it's descriptive or explanatory → RAG mode
- SQL Pipeline (Analytical): LLM generates SQL and executes → result is merged with context for a business-friendly answer.
- RAG Pipeline (Descriptive): LLM uses retrieved context to produce a clear, natural explanation.
- Hybrid Answer: For SQL questions, the system blends SQL output + semantic context + reasoning into one concise, insightful response.

References

- https://arxiv.org/abs/2005.11401v4
- https://www.researchgate.net/publication/388722115_Advancing_Retrieval-Augmented_Generation_RAG_Innovations_Challenges_and_the_Future_of_AI_Reasoning
- https://arxiv.org/html/2408.04948v1

Repository

- https://github.com/M22AIE241/MTP
- Execution Snippets: https://github.com/M22AIE241/MTP/tree/main/Validate_Execution

Future Scope:

- Fine-Tuning a Local SLM on Specific Data: Train a small language model (SLM) locally using knowledge distillation and LoRA to improve SQL generation accuracy, Domain-specific reasoning, Context understanding.
- **Query Planning & SQL Debugging Agent:** Introduce an LLM-powered “SQL Reviewer” that validates, corrects, and optimizes generated SQL before execution.
- **Multi-Modal RAG Integration:** Incorporate PDFs, images, dashboards, and logs into the vector DB for richer business insights beyond CSVs.