Data Warehouse Generator

Team: Abdelrahman Abdelnasser Gamal, Abdelrahman Adel Atta, Ahmed Reda Mohamed, Ahmed Mahmoud Mohamed, Arwa Amr Elsharawy, Alaa Emad Abdelsalam
Supervisors: Dr.Yasmine M. Afify, TA. Yasmine Shabaan
Information Systems Department, Faculty of Computer and Information Sciences - Ain Shams University

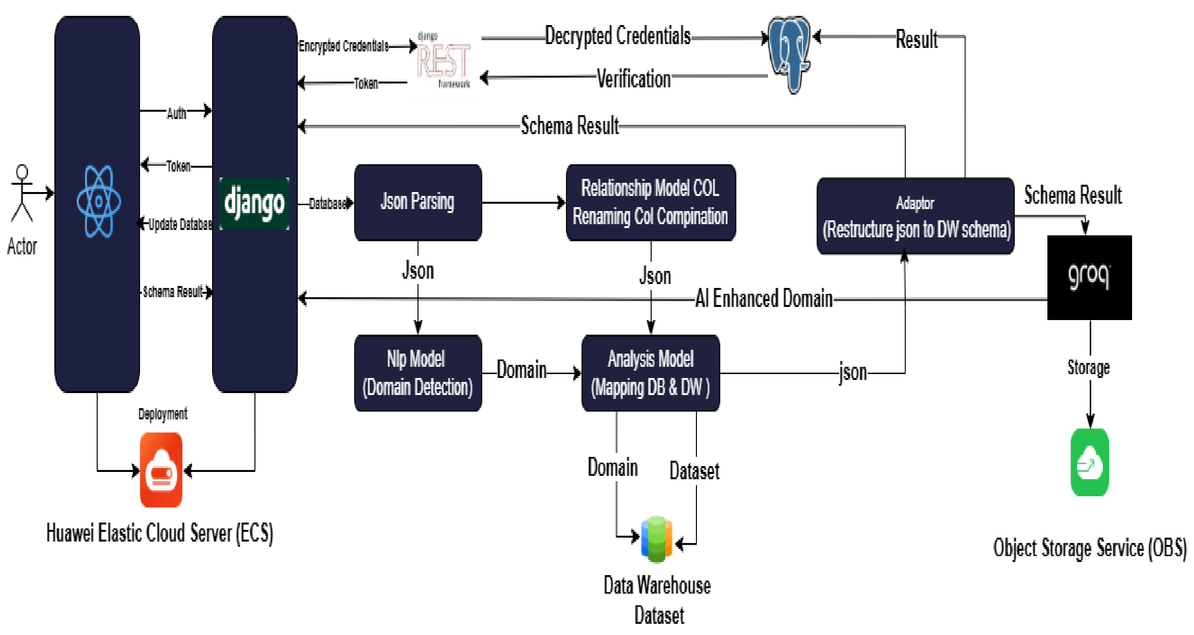
Introduction

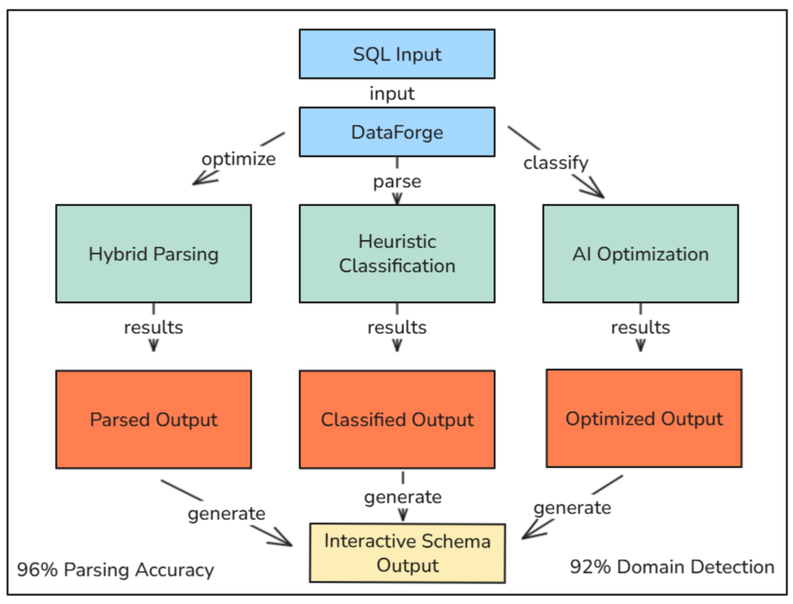
Do organizations face challenges when designing data warehouse schemas manually, especially in the era of big data?
Indeed, Manual processes are time-consuming, error-prone, and difficult to scale—leading to delays and increased costs. Studies show that 60% of organizations face project slowdowns, while 85% view faster analytics delivery as a strategic priority.
How can these challenges be addressed? DataForge introduces an automated approach, leveraging AI, heuristic classification, and hybrid parsing to streamline schema generation, minimize errors, and accelerate delivery.
What are the broader implications? This poster explores how **DataForge** enhances efficiency and supports data-driven decision-making across key sectors such as finance, healthcare, and retail.

Methods

How can we systematically extract and optimize data warehouse schemas to meet modern demands? DataForge employs a multi-stage, sophisticated process to achieve this: SQL Parsing and Analysis: Utilizes a hybrid approach combining regex patterns and Abstract Syntax Trees (AST) to extract table structures, columns, data types, and constraints from SQL DDL. This ensures compatibility across dialects like PostgreSQL and MySQL, handling complex nested constraints.
What advantages might this dual-method strategy offer over relying on a single parsing technique? Heuristic Schema Classification: Identifies fact and dimension tables by analyzing foreign-key density, numeric-column ratios, and cardinality thresholds, forming the basis for star or snowflake schemas.
How does calibrating these heuristics against expert-designed schemas enhance classification accuracy? AI-Driven Optimization: Leverages TF-IDF for keyword-based domain detection, BERT embeddings for semantic similarity, and LLMs (e.g., Google Gemini Flash) to suggest enhancements like audit fields or surrogate keys. A dedicated subsection on AI training reveals that the BERT model was fine-tuned on a corpus of 50,000 JSON-converted schemas, achieving 92.4% accuracy.
Why might domain-specific training data be critical for effective AI suggestions? Interactive Visualization: Renders schemas as draggable, color-coded graphs using React and ReactFlow, enabling real-time editing and validation.

System Architecture





DataForge Workflow: Automated Schema Generation Process

Key Algorithms

- Regex-Based SQL Parsing
- BERT Fine-tuned Domain Detection
- Vector Similarity Matching
- Rule-Based Enhancement Engine
- Gemini Schema Validation & Data Warehouse Enhancement
- Advanced Tokenization & Realistic Data Generation

Results

DataForge excels across diverse datasets (AdventureWorksDW, ShopSmart, TPC-DS), achieving 96% parsing accuracy for complex SQL DDL and 92% domain detection accuracy, adapting to contexts like retail and healthcare. It generates 100-table schemas in under 5 seconds (e.g., 4.0s for ShopSmart), far surpassing manual methods. Interactive visualization renders 50-node schemas in 2 seconds, ensuring seamless user interaction. User testing (n=10) yields a 4.2/5.0 satisfaction score, with 90% praising real-time validation and editing. Error reduction reaches 80%, AI enhancements improve schema quality by 5.1%, and compliance with dimensional modeling best practices rises to 75.6%. **DataForge** supports multiple SQL dialects and ensures robust referential integrity, enabling scalable, reliable data warehousing for efficient analytics.

Performance Metrics of Proposed Solutions

CHALLENGE	SOLUTION	OUTCOME
SQL Dialect Variance	Hybrid AST/regex + pre-normalization	96 % parsing accuracy
Domain Misclassification	Metadata enrichment + BERT fine-tuning (92.4 % acc)	92 % detection accuracy
Large-Schema Scalability	Batch FK analysis, caching, parallelization	4 s generation time
Visualization Lag	Lazy-load nodes, cluster UI	2 s rendering time
Invalid Edits	Real-time validation + undo feature	80 % error reduction

Dataset Performance Summary

DATASET	PARSING ACCURACY (%)	DOMAIN DETECTION (%)	GENERATION TIME (S)	VISUALIZATION LOAD (S)	USER SATISFACTION (1–5)	ERROR REDUCTION (%)
AdventureWorksDW	96%	92%	4.0	2.0	4.2	80%
ShopSmart	96%	92%	4.0	2.0	4.2	80%
TPC-DS	95%	90%	4.5	2.3	4.5	75%

Conclusion

DataForge addresses manual schema design errors and delays using AI-driven parsing, heuristic classification, and interactive visualization. Leveraging NLP with BERT and LLMs like Google Gemini Flash, it achieves 96% parsing accuracy and 92% domain detection, reducing errors by 80% and design time to under 5 seconds for 100-table schemas. The React-based interface ensures scalable, error-resistant warehousing, enhancing analytics efficiency across sectors.

Bibliography

1.Cormier, K., Zhang, K., Padron-Uy, J., Wong, A., Gagnier, K., & Parihar, A. (2025). Data warehouse design for multiple source forest inventory management and image processing.

2.Belhassen, Z., & Tlili, M. A. (2025). A novel framework for RDF schema extraction in NoSQL databases using Sentence-BERT.

3.Google Cloud. (2025, May 16). Techniques for improving text-to-SQL. Google Cloud Blog.