# Data Warehouse Generator

**Team:** Abdelrahman Abdelnasser Gamal, Abdelrahman Adel Atta, Ahmed Reda Mohamed, Ahmed Mahmoud Mohamed, Arwa Amr Mohammed, Alaa Emad Abdelsalam

**Supervisors :** Prof. Dr. Yasmine Afify, TA. Yasmine Shabaan

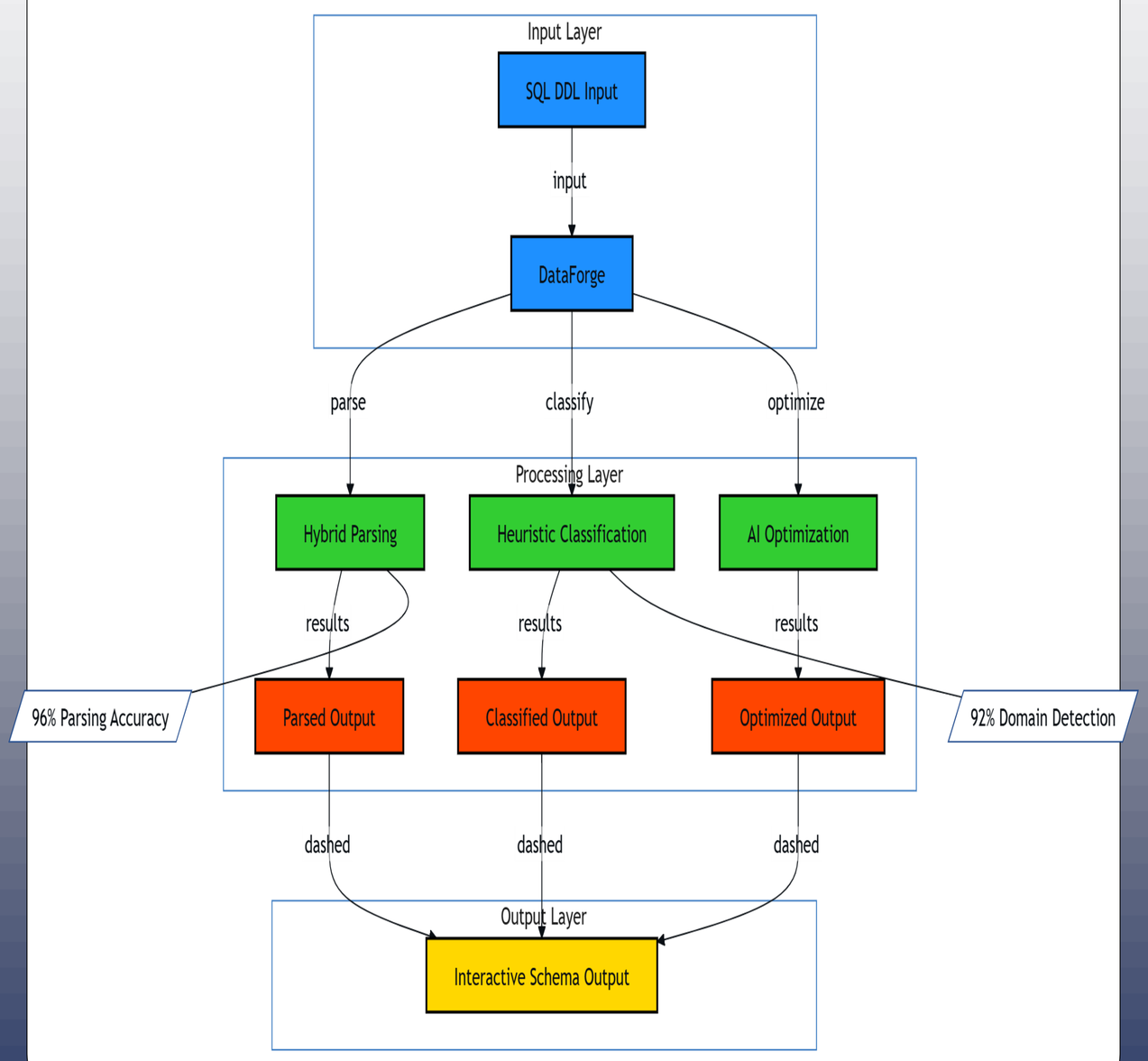
Information Systems Department, *Faculty of Computer and Information Sciences - Ain Shams University*

## Introduction

What challenges do organizations face in the era of big data when designing data warehouse schemas manually? The process is notoriously time-intensive, often spanning weeks, and prone to errors such as missing keys or inconsistent naming conventions, which can delay analytics and inflate costs. Industry trends show that 60% of organizations experience delays due to these bottlenecks, with 85% citing faster analytics delivery as a competitive necessity. DataForge, a groundbreaking tool developed by a team from Ain Shams University, tackles these issues by automating schema creation with AI and heuristics. How might this automation shift the paradigm from labor-intensive design to efficient, error-resistant workflows? This poster delves into DataForge's innovative approach, integrating hybrid parsing, heuristic classification, and AI-driven enhancements to revolutionize data warehouse design, ultimately empowering data-driven decision-making across diverse sectors like retail, healthcare, and finance.

## 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. How could this interactivity empower users to customize schemas according to specific business needs?




## 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%. The system supports multiple SQL dialects and ensures robust referential integrity, enabling scalable, reliable data warehousing for efficient analytics.

| CHALLENGE | SOLUTION | OUTCOME |
|---|---|---|
| SQL Dialect Variance (3.3.1) | Hybrid AST/regex + pre-normalization | 96 % parsing accuracy |
| Domain Misclassification (3.3.2) | Metadata enrichment + BERT fine-tuning (92.4 % acc) | 92 % detection accuracy |
| Large-Schema Scalability (3.3.3) | Batch FK analysis, caching, parallelization | 4 s generation time |
| Visualization Lag (3.3.4) | Lazy-load nodes, cluster UI | 2 s rendering time |
| Invalid Edits (3.3.5) | Real-time validation + undo feature | 80 % error reduction |

| 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 |

## Conclusions

DataForge automates data warehouse schema design with AI-driven parsing, heuristic classification, and interactive visualization, significantly reducing design time while ensuring high accuracy. Its intuitive interface and domain adaptability empower organizations to create reliable, optimized schemas, enhancing analytics efficiency and decision-making. By integrating advanced NLP and real-time validation, DataForge streamlines workflows and supports scalable, error-resistant data warehousing across diverse sectors.

## Bibliography

1. DiScala, M., & Abadi, D. J. (2016). Automatic generation of normalized relational schemas from nested key-value data. *SIGMOD*. http://www.cs.umd.edu/~abadi/papers/schemagen-sigmod16.pdf
2. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://arxiv.org/abs/1810.04805
3. Dwivedi, V. P., Jaladi, S., Fey, M., & Leskovec, J. (2025). Relational graph transformer. *arXiv*. https://arxiv.org/abs/2505.10960