White Paper: SAS to Snowflake SQL & Snowpark Converter Using Generative AI

Executive Summary

In today's data-driven landscape, migrating legacy analytics workloads from SAS to modern, cloud-native platforms like Snowflake and PySpark (Snowpark) is a strategic necessity. However, manual translation of SAS logic is time-consuming, error-prone, and often requires deep cross-technology expertise.

To bridge this gap, we present a Streamlit-based Generative AI-powered application that enables seamless, automated conversion of SAS code to both Snowflake SQL and Snowpark PySpark. By leveraging the Gemini API, this solution dramatically reduces conversion time, increases accuracy, and supports rapid modernization initiatives.

Architecture Overview

Frontend:
- Built with Streamlit for an intuitive, lightweight web interface.
- Allows users to upload `.sas` files or paste SAS code directly.
- Provides dual-tab interface for selecting either Snowflake SQL or Snowpark PySpark as output formats.

Backend:
- Utilizes Google’s Gemini 1.5 Flash model through its Python SDK.
- Dynamically prompts the model with context-aware instructions for precise translation.
- Applies optional static and live syntax validation to ensure output quality.

Key Features

- Dual Mode Input: File upload or direct SAS code input.
- Snowflake SQL Conversion: Clean, executable SQL.
- Snowpark PySpark Output: Compatible PySpark using Snowpark API.
- Static Syntax Validator: Checks for basic SQL correctness.
- Optional Live SQL Validation using Snowflake.
- Export: Download converted `.sql` or `.py` code.

Prompt Engineering Approach

Prompts are designed with:
- Clear role instruction: 'You are an expert in translating SAS to Snowflake SQL'
- Zero-shot directive to omit comments and produce executable code only
- Contextual language to guide formatting and eliminate hallucinations

Security & DevOps

- API keys and credentials are stored securely using dotenv.
- Can be deployed on Streamlit Cloud, Hugging Face Spaces, or internal infrastructure.
- Validation and logging modules can be added.

Benefits

- Speed & Efficiency: Convert code in seconds.
- Accuracy & Reliability: Verified through AI + validation.
- Developer Friendly: Easy-to-use UI.
- Cloud Native: Built for Snowflake and PySpark.

Future Enhancements

- Multi-agent comparison (e.g., Gemini vs OpenAI)
- RAG for schema-aware generation
- Full SAS to dbt conversion
- Visual diff tools
- Exportable logs and history

Conclusion

This tool represents a major leap in AI-driven code migration. With an easy UI and high-quality output, it empowers enterprises to modernize their analytics stack confidently.