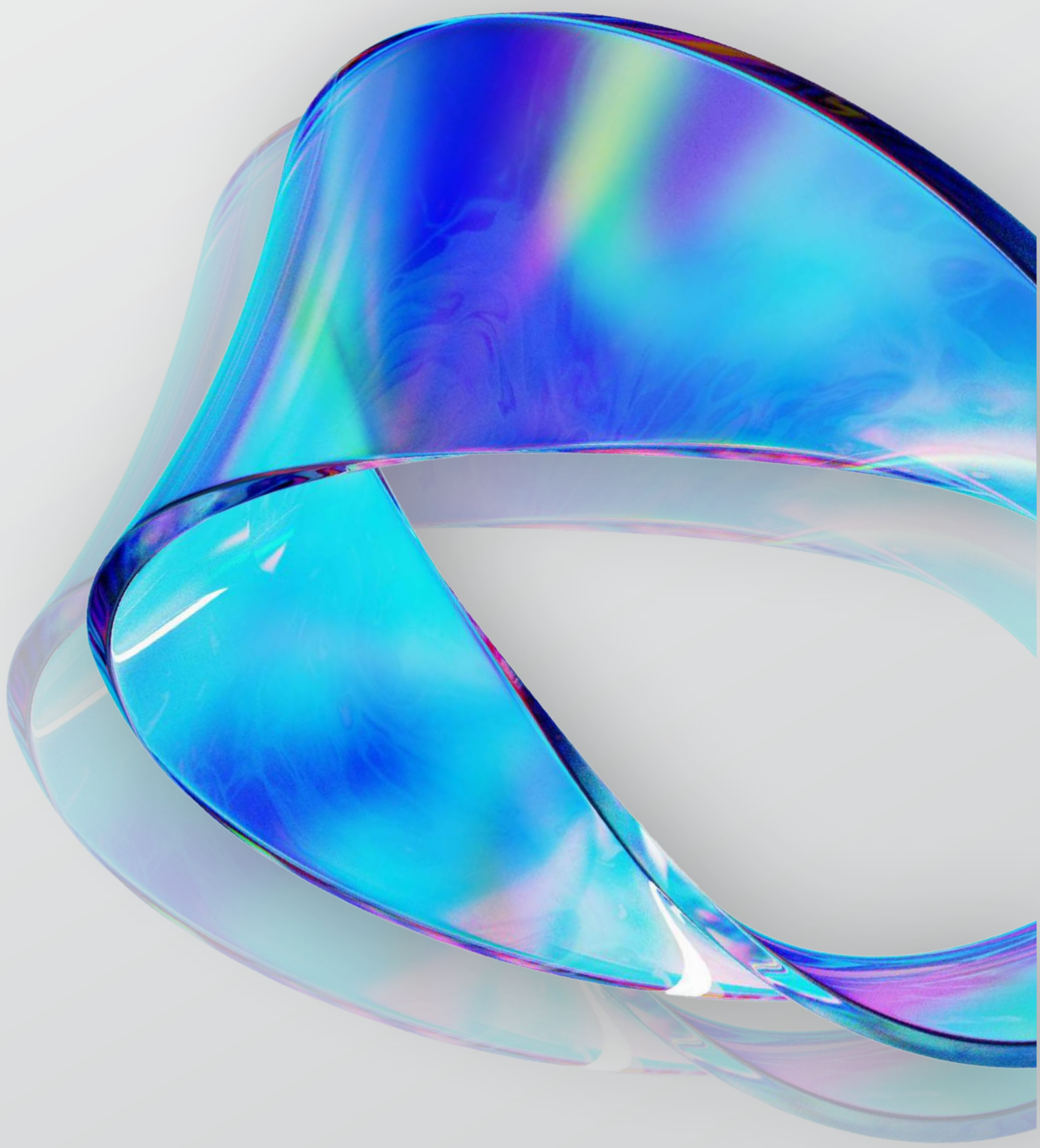
Kumo
Snowflake Native App

Architecture White Paper

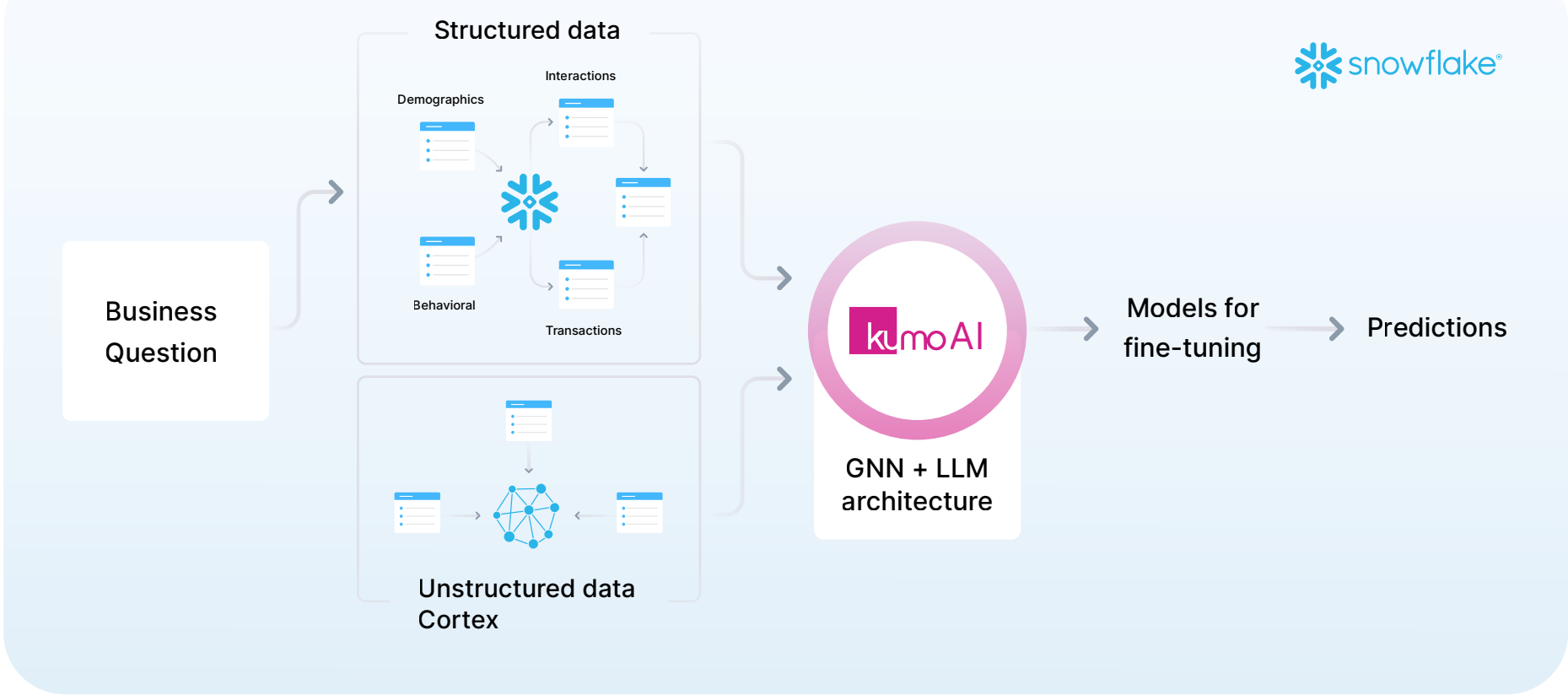


Introduction

Kumo offers predictive AI at scale by simplifying the end-to-end process and accelerating time-to-value. Central to this transformative approach is an AI purpose-built for finding hidden patterns in relational data by using both large language models (LLMs) and graph neural networks (GNNs) in a combination called GNN-LLM. Kumo uses the interconnected relationships in such data to make highly accurate predictions about behaviors, segments, lifetime value, and more to improve revenue impacting KPIs. By applying deep representation learning to relational data, Kumo eliminates the need for feature engineering, speeding-up the machine learning workflow and leading to more accurate models.

Users specify learning tasks in a declarative SQL-like **predictive query** language. Kumo executes the query and automates the entire process of data preparation, label engineering, training dataset creation, model optimization, and MLOps, simplifying the user's task of building a high performance model. Kumo's models can be further refined by domain experts for maximum performance. In hours, Kumo generates batch predictions or embeddings for downstream use with very little code.

Kumo in Snowflake (Private Preview) shortens the entire process, making it as easy to query the future with predictive AI as it has been to query the past with SQL – **all without the data ever leaving the Snowflake environment.**



Benefits of using Kumo in Snowflake

Kumo runs in Snowflake using Snowpark Container Services, a fully managed container offering designed to facilitate the deployment, management, and scaling of applications like Kumo within the Snowflake ecosystem.

This approach allows complex containerized applications to be easily and securely shared with Snowflake customers, providing them complete control of the application.

Retain Snowflake's security, compliance, and governance

Because Kumo runs inside the client's Snowflake environment, it inherits all Snowflake security, governance rules, and capabilities like column and row level security. Kumo seamlessly integrates with existing Snowflake workflows and policies for managing data by directly using Snowflake as the data storage and processing platform. Use Snowflake Single Sign-On (SSO) for authentication.

Control all data inside Snowflake

When using Kumo in Snowflake, no data leaves the client's Snowflake environment. Data and artifacts generated by Kumo are stored securely within the client's Snowflake environment as new tables of predictions or embeddings. These tables comply with Snowflake client data retention policies and are never accessed or managed by Kumo. Once written, the predictions and embeddings can be leveraged directly by other services for use in business operations.

Manage access: The client retains complete control over access to Kumo using Snowflake's access control mechanisms for the duration of the deployment. Clients must explicitly grant permission to any objects it wishes Kumo to access, which can be revoked or revised at any point in time.

Build on top of Snowflake

Kumo's seamless integration with the Snowflake data platform means that existing Snowflake workloads like ETL and EDA and code can be reused.

Improve Model Performance

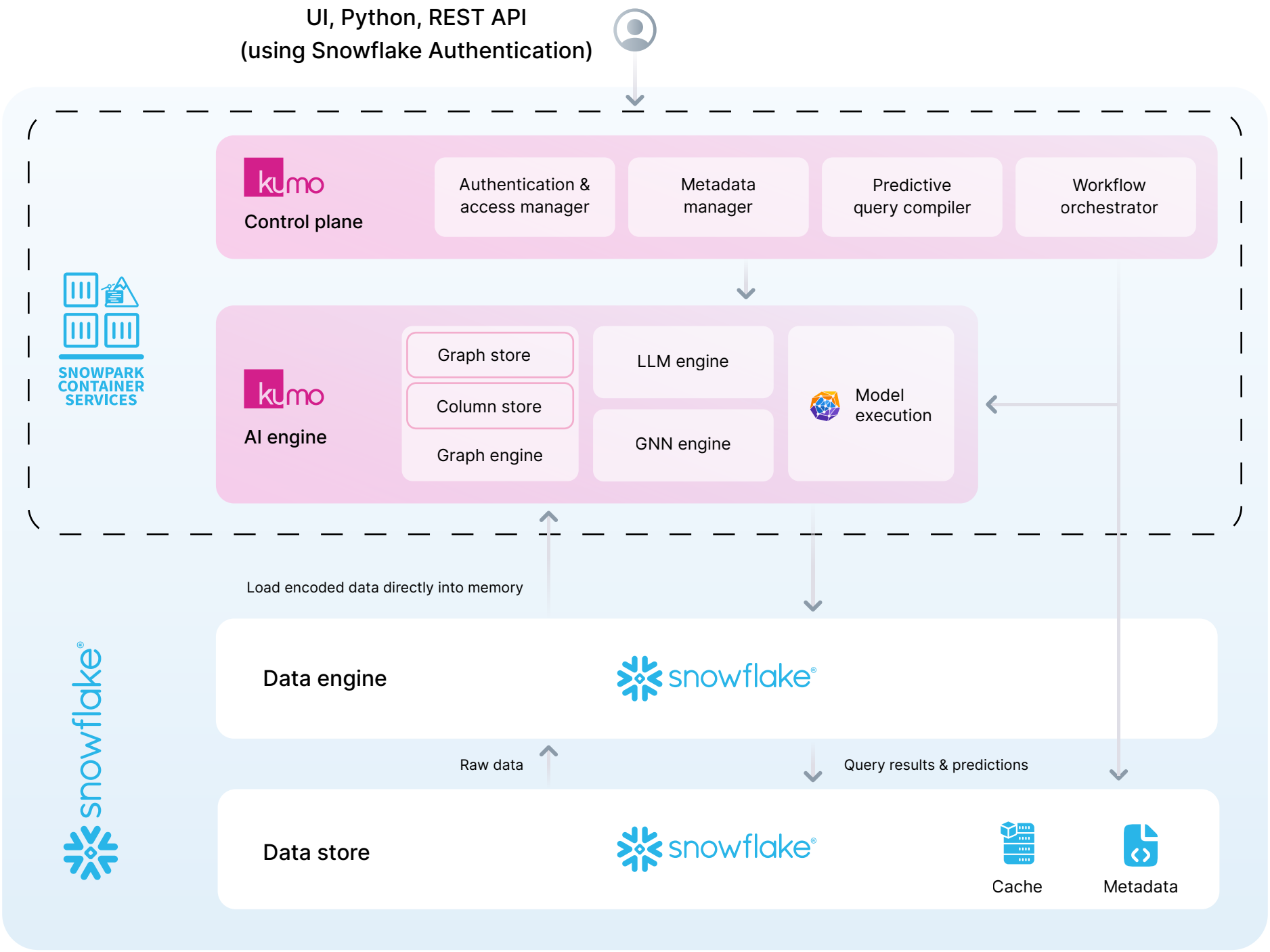
Kumo improves accuracy by eliminating hand-crafted feature engineering and instead learning embeddings directly from raw relational data, while providing advanced controls so the data scientist can add business value at each step.

Answer business questions in hours - not months

Kumo learns directly from your Snowflake tables, eliminates time-consuming feature engineering and instead uses AI to generate a model in hours, not months. Kumo's automated pipelines keep models fresh so the AI always learns from the latest Snowflake tables.

Architecture

Kumo is available in Snowflake using Snowpark Container Services. Kumo runs multiple containers (UI, REST, workers and trainer) that perform various functions to allow users to connect, process data and build learning models needed for predictive analytics.



Kumo's architecture consists of two key layers:

Control Plane

The control plane is a collection of services that include (a) an authentication and access control manager that integrates with client's Snowflake access controls, (b) a metadata manager that manages all metadata in Kumo, (c) a compiler that both translates predictive queries to an execution plan for the graph model and is used to generate predictions, and (d) a workflow orchestrator to coordinate various activities across Kumo.

Kumo AI Engine

The **Kumo AI engine** is a heterogeneous distributed system that processes predictive queries and builds predictive models. It contains multiple components, each responsible for a specific function in the ML pipeline.

The AI engine uses Snowflake as the Data engine. It leverages the Snowpark API to process input relational data in a client's Snowflake Data Cloud, and generates graph and training data that is used subsequently to build Kumo's AI model (GNN-LLM). The metadata and data for these outputs are materialized to a Snowflake stage, which acts as a cache for the data, in the client's Snowflake account.

The **Graph engine** loads graphs and (node) attributes generated by the Data engine, in the Graph and Column store respectively. Its primary role is to serve subgraph sampling requests and node attributes for the GNN training and batch inference.

- The **Graph Store** contains all entities in the data warehouse (rows).
- The **Column Store** contains all the attributes about the entities (columns).

The **GNN engine** is responsible for GNN computations and batch inference on the relational data.

The **LLM engine** is responsible for working with the pretrained LLM, which brings general-world knowledge to customer's relational data via Kumo's GNN-LLM. Kumo works with both open-source and commercial LLMs.

Model execution combines the GNN and the LLM parts of the architecture and uses the available GPU compute in Snowflake to train the GNN-LLM model for a given predictive query on customer's data.

Kumo's **GNN-LLM models** learn from graph-structured data and use the leading open-source framework, PyTorch Geometric, for **model execution**, built and maintained by Kumo members. Kumo leverages a variety of graph neural network architectures and training procedures especially designed for learning on relational databases.