

KANO TM Pending

Overview

KANO at a glance

Gen AI | Multi-modal | Agentic | GraphRAG

PLATFORM

KANO is a generative AI solution for complex knowledge management challenges where 'traditional RAG' does not deliver

DATA

KANO unlocks insights over *multiple sources*

Structured **text**





Unstructured text

Diagrams which embed crucial information



EXPLOITATION

KANO can be leveraged through *multiple channels e.g.*

-  Chat
-  Document generation
-  Exchange with other knowledge tools
-  Smart agents

可能 (Kanō) means 'possibility' in Japanese, which reflects both the types of scenarios we can address and the insights we can make possible which were previously unachievable without significant human effort

KANO creates a *knowledge graph*



You will have seen a knowledge graph if you've ever watched a crime drama on TV!

Where there are many pieces of evidence with different relationships in time and space, a knowledge graph brings them together.

In modern policing, graph databases replace string and pins – although they don't make for good TV!

Knowledge graphs are powerful but until the advent of Gen AI, it was hard to populate them from unstructured knowledge.

Academic research (including from Microsoft) highlights the benefits of using graphs and Gen AI together; KANO is an early commercial implementation.

What makes KANO different?

Gen AI | Multi-modal | Agentic | GraphRAG

GRAPHS INSIDE




The knowledge graph is at the heart of every KANO deployment; almost any kind of declarative knowledge can be stored in this way (see below), making the solution flexible

GRAPHS EVERYWHERE

KANO recognises that the most valuable enterprise knowledge is highly interconnected; rich answers must draw from closely connected concepts

GRAPHS REFINED

KANO uses a range of techniques to:

-  Maximise the size & rich interconnections of the graph
-  Optimise graph traversal for queries
-  Maximise the fidelity of the graph








With every vendor and system integrator offering gen AI capabilities, KANO does not target commodity solutions, which represent the majority of offerings

What use cases need KANO?

Gen AI | Multi-modal | Agentic | GraphRAG

Simple cases





There are many Generative AI Retrieval Augmented Generation (RAG) solutions on the market and integrated into products; they give strong ROI for use knowledge use cases where:

-  The source information is short (e.g. an email) or self-contained (e.g. a corporate policy, a meeting transcript or product manual)
-  Input information is text-only or contains simple pictures such as flowcharts
-  Is repetitious but not interconnected ('summarise the top complaints')
-  The outputs are straightforward ('explain how feature X works')
-  Language translation is required

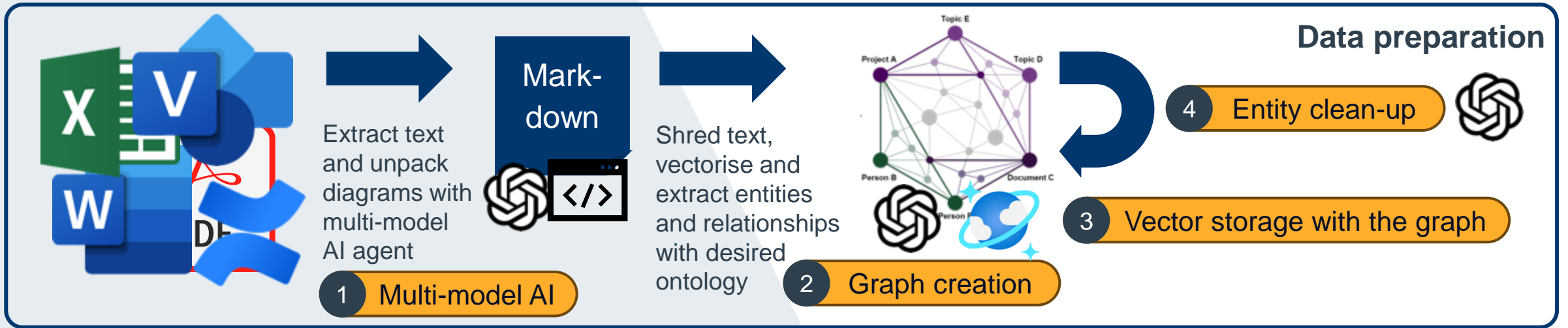
vs.

Wicked problems

KANO can *cover the simple cases* and excels in all the following scenarios:

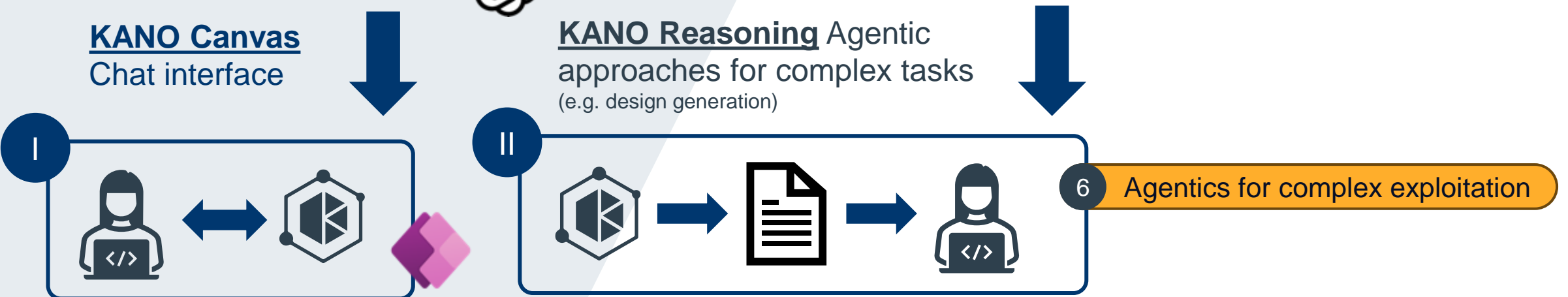
-  Concepts are highly interconnected across the knowledge sources (e.g. IT landscapes, complex applications, physical infrastructure, supply chains, legal cases, biomedicine)
-  Some knowledge is in the the form of diagrams which contain meaning not available elsewhere (and therefore must be contextually broken down to unlock knowledge)
-  Fine-grained access control is needed over knowledge sources (since each element of the graph can be secured);
-  Structured and highly repeatable information is needed in the answer (e.g. to generate diagrams) or knowing what information relates to what timeframe (historisation) is crucial

Under the hood: KANO brings together 6 powerful features



5 **Choose how to use the graph depending on query**

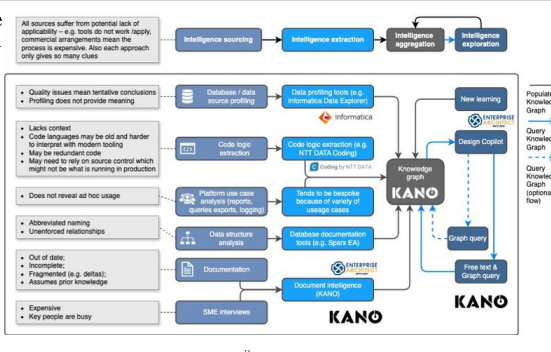
Graph + vector search combined



Use cases tackled so far

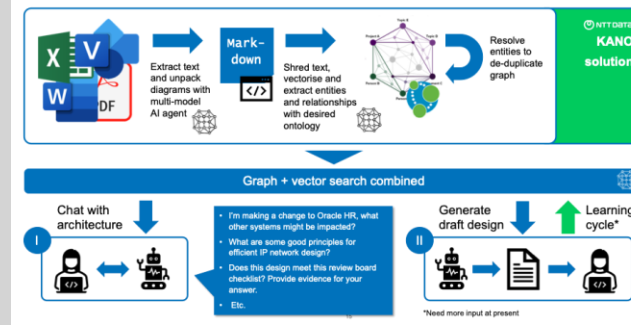
As we work with customers, we are finding more use cases

Possible solution



Unpacking complex applications through code & documentation; onboarding staff

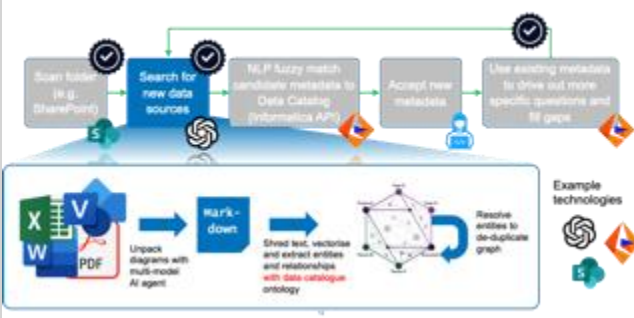
What we built



Understanding IT landscape and generating new design documentation

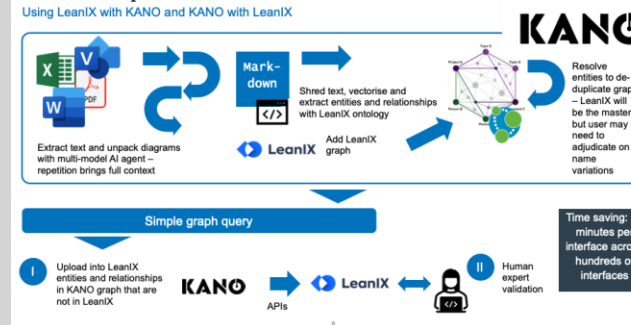


Applying to data cataloguing domain



Data landscape discovery

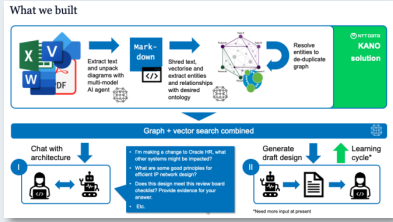
LeanIX concept



Populating Application Portfolio Management (APM) tools

Business case for KANO

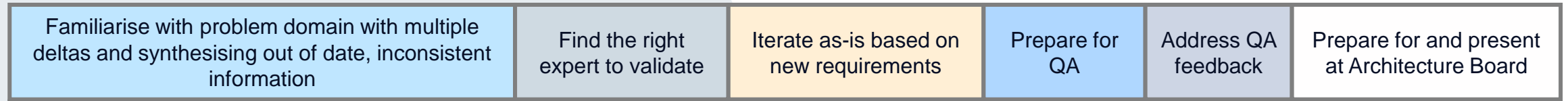
We have explored the business case for one use case: “Understanding IT landscape and generating new design documentation”



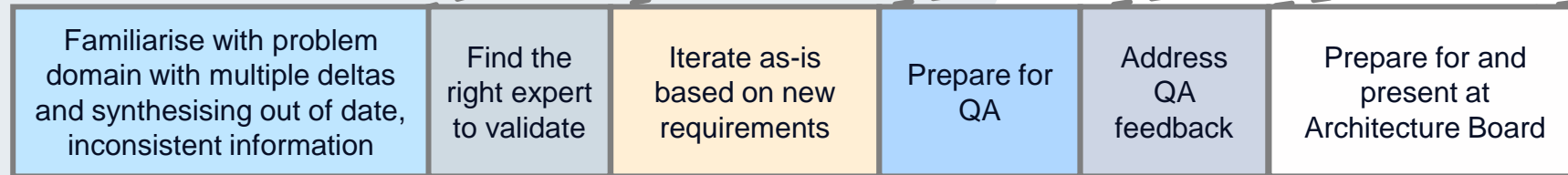
Life of the architect...



As-is



To be



Rationale with KANO

Faster information gathering – acronyms expanded, inconsistencies highlighted, etc.

Explore the social network of experts

Gen AI creates the first draft based on best practice

Summarise and consume information more rapidly as a reviewer; solution learns from human feedback

Re-format information for boards; auto-QA against board criteria


Estimate at least **20% saving** – less up-front effort; fewer re-work cycles.

Getting started

We can run a 4–5-week POC where clients typically see value with their own data



Setup

- KANO is ready to deploy on Azure 
- We deploy into your cloud and within your security boundaries
- We require access to a foundation model with multi-modal capabilities (via OpenAI)
- We manage graph database licenses on your behalf



What we need from you

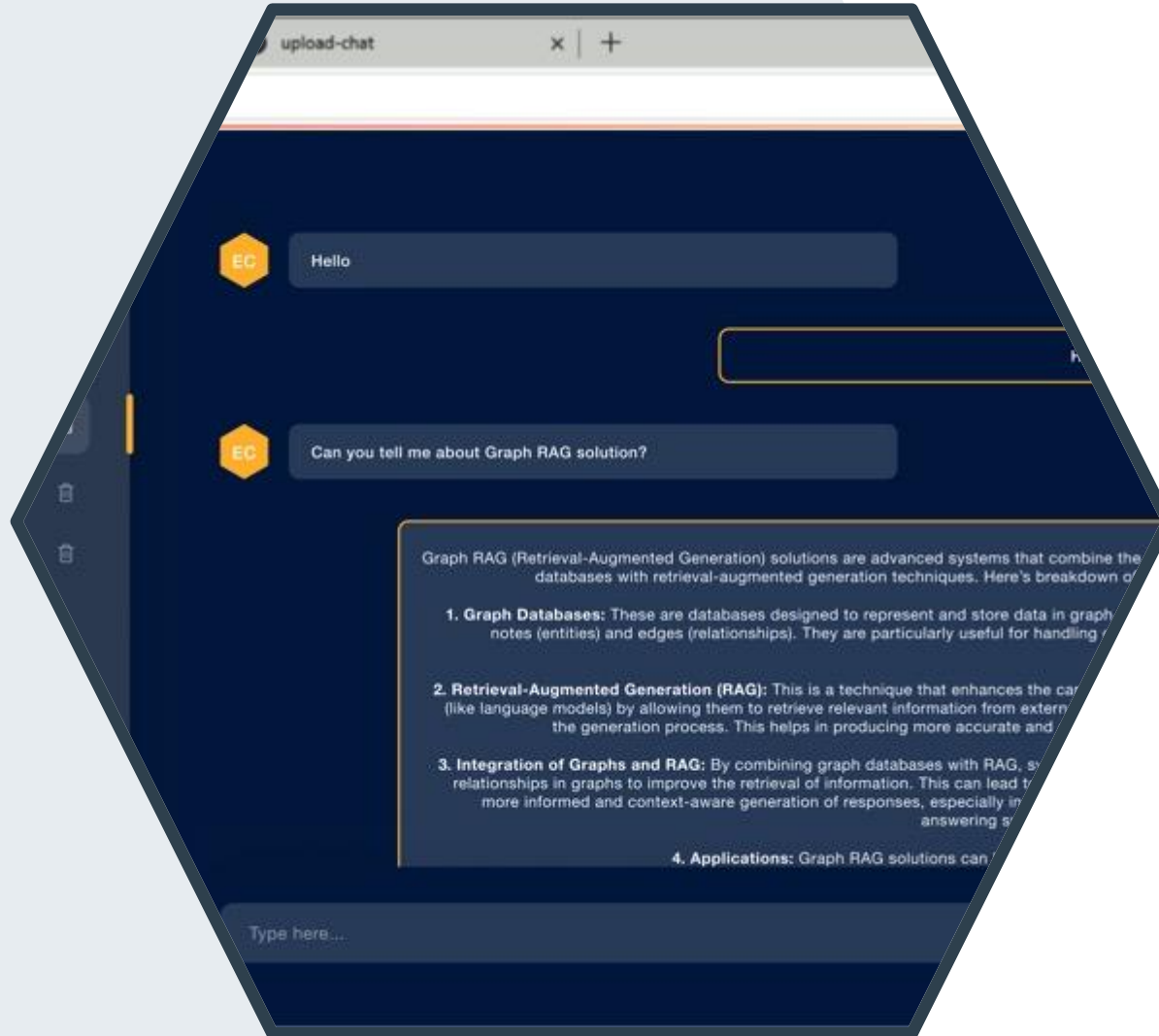
- While KANO can use a number of data sources, we find that a selection of documents is a good starting place.
- Like any Generative AI solution, the results are not always 100% accurate, so we work with your teams to fine-tune the retrieval and understand the level of efficiency gain. We need access to your experts for this to work.



Bear in mind

- KANO is cutting-edge technology; it is deployed as a set of foundation libraries that are deployed and customised for each client. It is not yet a shrink-wrapped product.
- Think of KANO as a clever but not infallible member of the team; it needs context for best results (e.g. to interpret diagrams)

Show don't tell



We have a demo using open-source data which can be shown to any customer. It is through the demo and dialogue that the powerful features of the platform become obvious.

In particular we can show RAG and GraphRAG side by side and demonstrate the generation of new design documentation.

Comparison: GraphRAG vs. RAG

Enter your question or prompt
I would like to deploy a graphRAG solution using Next.js as the vector and graph store into AWS, can you provide me with an architecture to deploy?

GraphRAG

Certainly! I'll provide a detailed architecture and explanation for deploying a GraphRAG solution in AWS using Next.js as both the vector and graph store, based on the prompt question. While the relevant information doesn't specifically address a Next.js-based solution, I'll adjust the examples to fit your requirements.

Architecture Diagram (Simplified representation):

```

graph TD
    User[User] --> API[API Gateway]
    API --> Lambda[Lambda Query Handler]
    Lambda --> S3[S3 Document Storage]
    S3 --> SQS[Ingestion Queue]
    SQS --> EC2[Vector Embedding Service]
    EC2 --> Secrets[Secrets Manager]
    Secrets --> IAM[IAM Roles & Policies]
    
```

Detailed Explanation:

- AWS Services Involved:**
 - Amazon API Gateway: Handles incoming user queries.
 - AWS Lambda: For query processing and data ingestion.
 - Amazon S3: Stores original documents.
 - Amazon SQS: Manages the ingestion queue.
 - Amazon EC2 or ECS: Hosts the vector embedding service.
 - AWS Secrets Manager: Stores sensitive information (e.g., Next.js credentials).
 - IAM: Manages roles and policies.
 - Amazon VPC: Provides network isolation.
 - AWS Direct Connect or VPN: For secure connection to Next.js (if using Next.js on-prem).
- Next.js Integration and Deployment:**
 - Deploy Next.js using Amazon ECS instances within a VPC for full control, or use Next.js Aurora (Next.js fully managed cloud service) on AWS.
 - Configure Next.js to support both graph operations and vector similarity search.
 - Use multi-AZ deployment for high availability.
- Data Flow - Ingestion:**
 - Documents are uploaded to S3.
 - S3 triggers a Lambda function.
 - Lambda function sends a message to SQS.
 - Another Lambda processes the SQS message, reads the document from S3.
 - The document is sent to the Vector Embedding Service (running on EC2/ECS).
 - Resulting vectors and metadata are stored in Next.js.
- Querying:**
 - User sends a query through API Gateway.
 - Lambda processes the query, generates a vector embedding.
 - Lambda queries Next.js for similar vectors and relevant graph data.
 - Results are processed and returned to the user.
- API Gateway and Load Balancing:**
 - API Gateway handles incoming user requests.
 - If using EC2 for Next.js, consider using an Application Load Balancer for distributing requests across Next.js instances on AWS.
 - Configure Next.js to support both graph operations and vector similarity search.
 - Use multi-AZ deployment for high availability.
- Data Flow - Querying:**
 - Documents are uploaded to S3.
 - S3 triggers a Lambda function.
 - Lambda function sends a message to SQS.
 - Another Lambda processes the SQS message, reads the document from S3.
 - The document is sent to the Vector Embedding Service (running on EC2/ECS).
 - Resulting vectors and metadata are stored in Next.js.

RAG

I apologize, but I don't have specific information about deploying a graphRAG solution using Next.js as the vector and graph store in AWS based on the content provided. The content describes a RAG-based application using Aurora PostgreSQL and Amazon Neptune, but it doesn't mention Next.js or a graphRAG solution.

If you need a precise architecture for your specific requirements, I would need more information about graphRAG implementations with Next.js in AWS. If you're looking for a similar architecture but with Next.js instead of Aurora PostgreSQL, you might need to adjust the graph architecture, but I can't provide specific details without more context.

If you need a precise architecture for this use case, I recommend consulting AWS documentation, Next.js resources, or seeking advice from a specialist in graph databases and AWS deployments.

RAG runs out of steam



Knowledge graph fragment

Case study | Telco

Augmented software architecture for 20% cost savings in design phase

KANO allows knowledge graphs to be extracted from enterprise documentation, transforming the understanding of legacy technology landscapes and speeding up the architecture and design lifecycle

Business need

VMO2 are one of the largest Telcos in the UK with 45 million customers and are relied on for critical national infrastructure.

They have a complex IT landscape documented in different ways with using different tooling. Many 'delta' design documents made it hard to understand as-is functionality. Inconsistencies had arisen over time and enterprise architecture tools were unevenly adopted.

When system changes were needed, a key challenge was the time taken for architects new to the client environment to get up speed. This negatively impacted time to market and imposed additional costs. Key members of staff with knowledge of core systems became over-stretched.

These are features of many large organizations and are classic knowledge management challenges. The problem had proven too complex for Gen AI solutions using RAG.

Solution

Our multi-cloud KANO solution addressed these challenges by:

- Using multi-modal AI to inspect and document technical diagrams which contained details not present in the text
- Extracting entities and relationships into a knowledge graph as well as vectorizing the text from underlying documents
- Offering a query interface to ask questions over the architecture knowledgebase using graph and vector approaches together (GraphRAG) for significantly higher accuracy and usefulness compared with using RAG alone (see example)
- Using agentic approaches to generate draft designs for architects (assisted by the query interface) to complete.

Collectively these techniques make KANO unique in the market at the time of release and suitable for many complex knowledge management challenges.

Outcomes

- ✓ **Faster onboarding:** Architects who are new to working with VMO2 can rapidly onboard and find the information they need
- ✓ **Faster QA:** In future releases senior architects can ask KANO to evaluate new designs with respect to existing standards
- ✓ **Faster software implementation:** The features of KANO can be used throughout the design phase, with an estimated saving in excess of 20% in time and effort
- ✓ **APM population:** The KANO knowledge base can be used to rejuvenate the enterprise architecture tools

Another 2 pages of relevant information...

Benefits of KANO (detail)



Better responses

- More robust way of **dealing with fragilities in source information**: avoids multi-hop inferencing that using RAG alone suffers with
- **Whole dataset reasoning**: *‘what are the most common interface patterns in our application estate?’*
- **Flexible**: we can choose when to use graph vs vectors vs both



But other useful features

- Excellent solution for when **exactly the same information must be used time after time** – e.g. for diagrams
- **Explainable**: easily see what structured information the solution ‘understands’ and easily make a few permanent corrections.
- **Interactive**: supports complex graph queries over enterprise metadata
- Forces some **standardisation**: how do we want to apply a knowledge ontology in each use case?
- **Integration-ready**: a solution that works with structured data and enterprise metadata you already have
- **Lower cost and emissions** where user needs can be satisfied by the graph alone
- By using Neo4J we can support **historisation** – for example understanding that some information will be relevant in a future architectural epochs.
- **Fine-grained access control** at the node or edge level

KANO is part of the wider NTT DATA UK Generative AI Story

Our portfolio of services



Literacy & learning

Foundational Literacy
Experiential GenAI
GenAI Syllabus



Strategy & understanding

GenAI Impact Assessment
GenAI Maturity Assessment & Strategy
Rapid Use case Identification (ShinrAI)
Rapid Prototyping Lab



Research and Design

AI Foresight backed by NT Group R&D in LLMs
Deep research and concept creation



Acceleration & Scaling

Solution Industrialisation
GenAI Product Factory
Implementation testing

Solutions and platforms



Operational Efficiencies

Augmented Architecture & Engineering (Axet suite + KANO)
GenAI platform development (TechHub)
Augmented Digital Workplace Services
GenAI BPS platform services for workforce delivery



Employee & Customer Experience Transformation

Knowledge Discovery Acceleration (KANO)
Ai Virtual Assistant (Syntphony)
AI-CX platforms for DWS & CX
Intelligent Document platforms (Syntphony)



AI-ready infrastructure

AI ready data-centre (hybrid/regulated)
AI-enabled networks infrastructure
Security services for AI



KANO

MASTERING COMPLEX KNOWLEDGE

For more information contact bill.m.wilson@nttdata.com

