modak ProcessTempo neo4j

Developing a Global Metadata Context Graph for Enhanced Data-driven Decisions

Why Develop a Global Metadata Context Graph?

In the contemporary data-driven landscape, organizations face challenges in accessing and utilizing data effectively. Siloed data, quality concerns, and accessibility issues hinder business users/ data and business teams from making informed decisions. While central data catalogs aim to address these challenges, they fall short in providing comprehensive context. This article explores the concept of a Global Metadata Context Graph (GMCG) to augment data catalogs, offering a practical guide for its development.

Challenges with Data Catalogs

Data catalogs help organize metadata, but challenges persist:

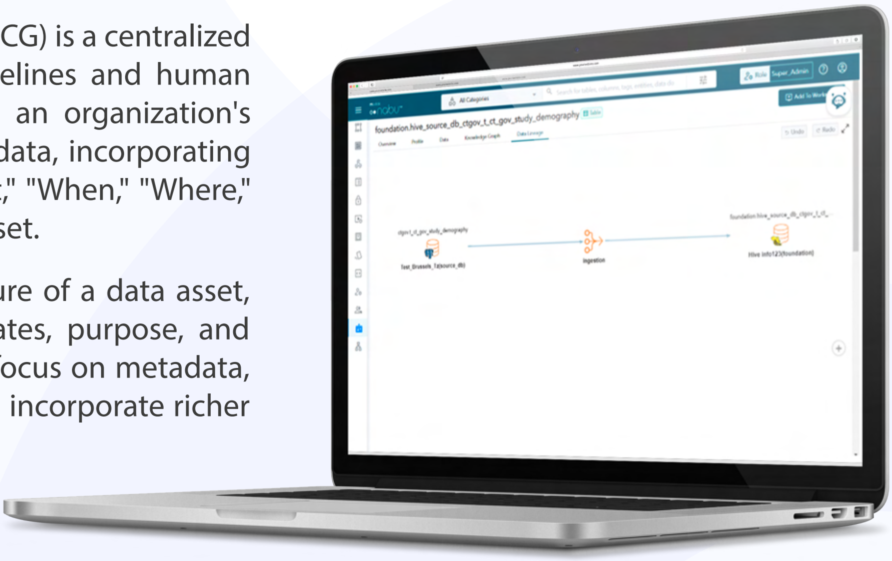
- Rapid changes in the data environment require continuous catalog maintenance.
- Multiple catalogs with varying technologies and maturity levels can exist within an organization.
- Data catalogs often lack comprehensive data quality indicators and fail to convey data freshness.
- Importance of Context in Data

Context is crucial for analysts seeking relevant data. A GMCG enhances data catalogs by providing additional context, answering questions about data lineage, usage, and freshness. This facilitates efficient decision--making and addresses the limitations of traditional catalogs.

What is a Global Metadata Context Graph (GMCG)?

A Global Metadata Context Graph (GMCG) is a centralized platform fed by automated data pipelines and human processes, enabling users to explore an organization's data landscape. It goes beyond metadata, incorporating context—answering the "Who," "What," "When," "Where," "Why," and "How" behind each data asset.

While metadata describes the structure of a data asset, context delves into its origins, updates, purpose, and usage. Traditional data catalog tools focus on metadata, necessitating the need for a GMCG to incorporate richer contextual information.


nabu
All Categories
Search for tables, columns, tags, entities, data do
Role Super_Admin
Add To Work
foundation.hive_source_db_ctgov_t_ct_gov_study_demography
Table
Overview
Profile
Data
Knowledge Graph
Data Lineage
Undo
Redo
ctgov.t_ct_gov_study_demography
Test_Brussels_T(source_db)
Ingestion
foundation.hive_source_db_ctgov_t_ct_...
Hive info123(foundation)

The Role of Graphs in GMCG

Graphs, as data constructs, store information to highlight relationships between various elements. In a GMCG, graphs facilitate easier exploration of data assets, offering benefits such as:

- Highlighting important information through relationships.
- Navigating information from different perspectives based on relationships.
- Outperforming traditional database structures, especially in complex data models.
- Reducing pressure on analysts to understand complex data models.

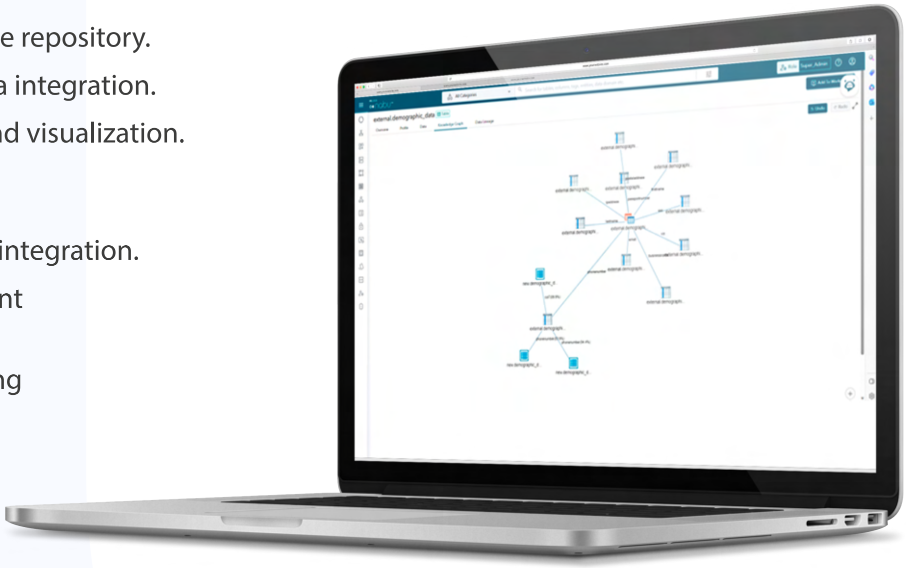
Graph databases like Neo4j are well-suited for GMCGs, providing a flexible and efficient core repository.

The Role of Graphs in GMCG

Developing a GMCG involves a combination of people, processes, and technology:

Technological Components:

- A graph database (e.g., Neo4j) as the core repository.
- An ETL/data pipeline for automated data integration.
- A UI/UX platform for front-end access and visualization.



Human Resources:

- Solution architect for system setup and integration.
- Data engineer for pipeline development and automation.
- Subject matter experts for data modeling and visualization.
- Product manager to coordinate efforts.

Processes and Workflow:

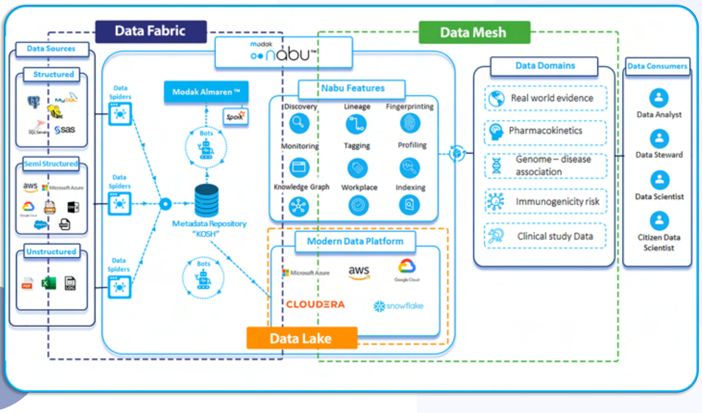
- Site administration for environment and user management.
- New asset process for introducing data into the environment.
- Change status, metadata enhancement, attestation, feedback, and exception handling processes.

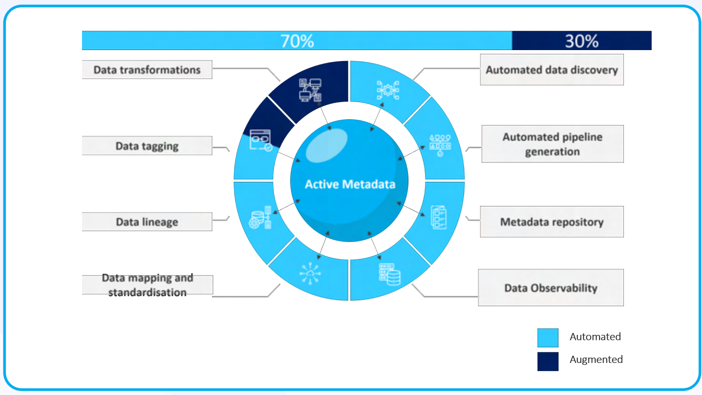
The data required:

A large organization may have several systems which capture important contextual information. These systems include:

- Data catalogs – a natural repository for metadata, these systems capture knowledge about the existence of data and allow users to enrich this data by identifying data stewards, or translating complex data terms into those that users may understand.
- ETL platforms – ETL systems capture a great deal of logging information as they process data. This information can provide very useful insight into the history and lineage of a given asset.
- Data platforms and repositories
- Business Intelligence (BI) tools platforms
- Internal systems and spreadsheets

Automating the collection of this data into a common data model is the first step in the development of the GMCG.





A Winning Combination

Why Neo4j?

Neo4j is chosen for its graph database capabilities, ideal for capturing and storing relationships, enabling users to understand the interconnectedness of tools and platforms. Its property graph model allows for faster iteration than rigid SQL-based approaches, facilitating quick implementation of global metadata catalogs, particularly in large, complex environments.

Why Modak?

Modak Nabu™, an integrated data orchestration platform, automates data preparation at a petabyte scale. It addresses the challenge of delivering trustworthy and contextual data for AI-driven initiatives, aligning with Gartner's recommendation of data fabric and data mesh architectures. Modak Nabu™ creates a Data Fabric from heterogeneous datasets, populating a Data Lake to form Data Domain products as per Data Mesh principles, serving various stakeholders and use cases.

Why Process Tempo?

Process Tempo, supporting Neo4j natively, accelerates graph solution implementation with its no-code, drag--and-drop platform. It combines data visualization and embedded workflow, allowing non-technical users to explore the graph without understanding complex query languages. As the front-end application for the central data catalog, Process Tempo ensures easy access, maintenance, and analysis, enhancing data and knowledge collection processes.