Graph Databases

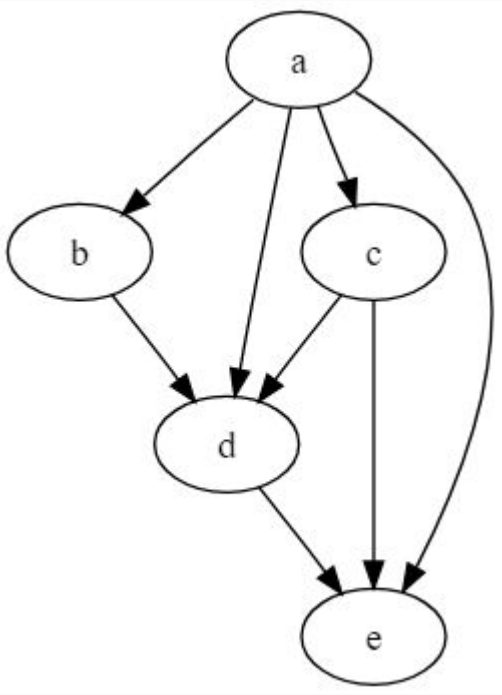
Group C2:
Zachary Breitbart
Mingxiao Song
Michel Maalouli

01 Introduction

Technology Overview · Development Trend · Features & Functions

Technology Overview

Graph databases are a type of NoSQL database that stores data in the form of graphs, emphasizing relationships (edges) between data points (nodes), allowing for efficient data modeling and querying.



File:tred-g.svg. Wikimedia Commons. (n.d.). https://commons.wikimedia.org/wiki/File:Tred-G.svg

Common Terms

Node

The main entities in a graph. Also known as vertices or points. Usually people, items, accounts. etc.

Edges

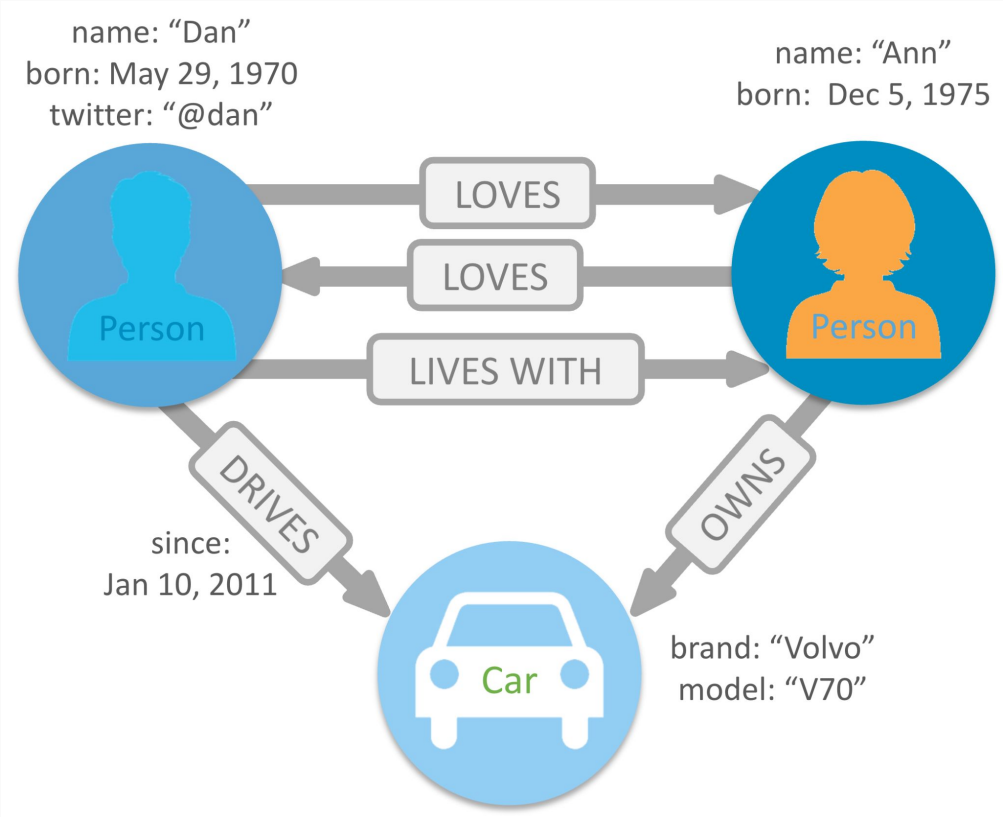
The connections between those entities. Also known as relationships

Label

Attributes that group similar nodes together;
Describes what a node is (Examples: car, or person).

Property

Key/value pairs containing information about nodes or edges (Example: person’s name or car’s make)

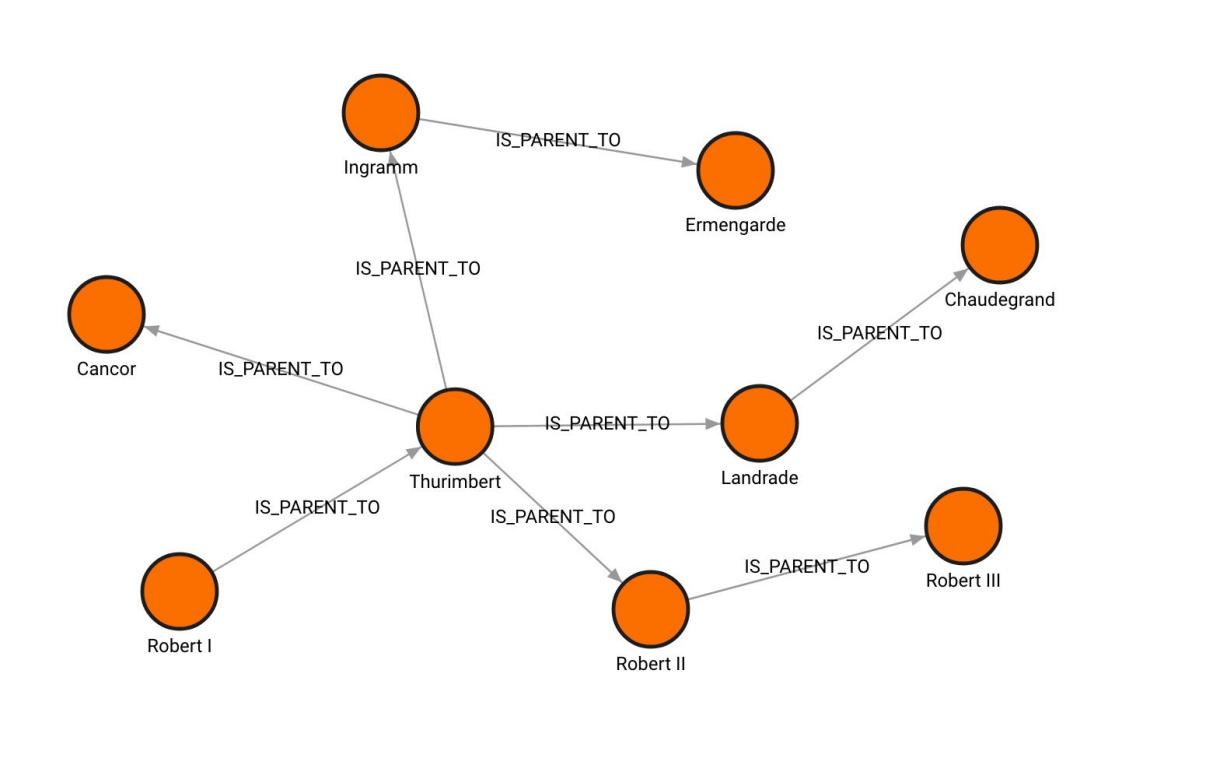


What is a graph database? - developer guides. Neo4j Graph Data Platform. (n.d.). https://neo4j.com/developer/graph-database/

Why Graphs?

- Better at representing data that has:
  - A large amount of connections that will be analyzed
  - Irregular and complex connections
  - Schemas that will change often

- Think of representing a family tree in a relational database
  - If we wanted to find the all the descendants of someone, we would have to do recursive joins which can explode query times
    - SQL syntax is not designed for this type of problem
    - One line when using the most popular graph query language, Cypher



Graph database vs relational database. Memgraph. (n.d.). https://memgraph.com/blog/graph-database-vs-relational-database

SQL

```
WITH RECURSIVE descendants AS
(
SELECT person
FROM tree
WHERE person='Thurimbert'
UNION ALL
SELECT t.person
FROM descendants d, tree t
WHERE t.parent=d.person
)
SELECT * FROM descendants;
```

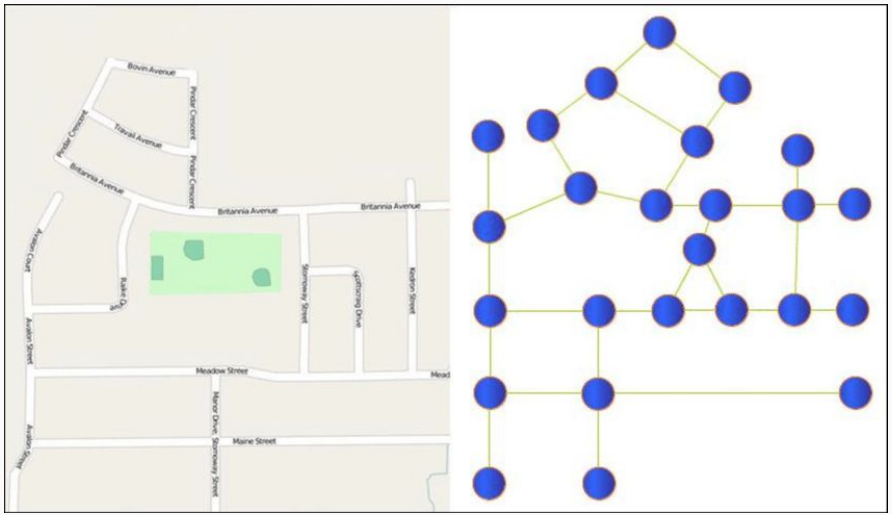
Cypher

```
MATCH path=(n:Person {name: 'Thurimbert'})-[*]->(m)
RETURN m;
```

Graph database vs relational database. Memgraph. (n.d.). https://memgraph.com/blog/graph-database-vs-relational-database

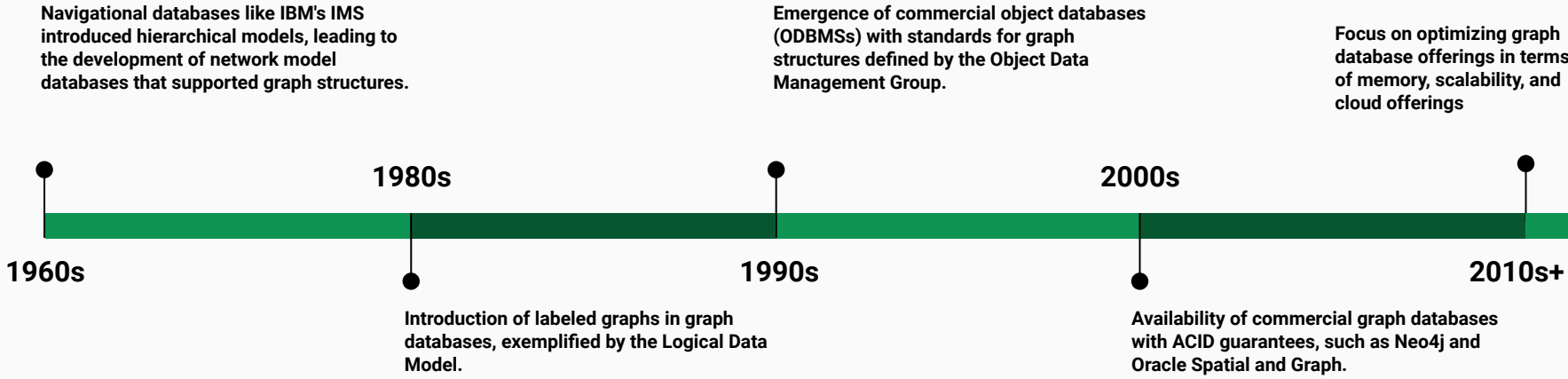
Business Applications

- Network Analysis
  - Social Media
  - Cellular Networks
  - IT Management
- Supply Chain Optimization
- Recommendation Systems
  - Advertising agencies
  - Movies, music, and games
- Infrastructure



SOCIAL-BASED TRUSTWORTHY DATA FORWARDING IN VEHICULAR DELAY TOLERANT NETWORKS - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/Figure-3-2-Representing-roads-map-into-a-graph_fig2_266497044 [accessed 29 Feb, 2024]

History


Navigational databases like IBM's IMS introduced hierarchical models, leading to the development of network model databases that supported graph structures.
Emergence of commercial object databases (ODBMSs) with standards for graph structures defined by the Object Data Management Group.
Focus on optimizing graph database offerings in terms of memory, scalability, and cloud offerings
1960s
1980s
1990s
2000s
2010s+
Introduction of labeled graphs in graph databases, exemplified by the Logical Data Model.
Availability of commercial graph databases with ACID guarantees, such as Neo4j and Oracle Spatial and Graph.

Typical Features

- **Index-Free Adjacency**: Fast retrieval of connected elements without indexes.
- **ACID Compliance**: Ensures reliable transaction processing.
- **Horizontal Scalability**: Manages workload of large datasets
- **Flexible Schema**: Dynamic addition of new elements.
- **Specialized Query Languages**: Simplify complex relational queries.

Typical Features

- **Graph Traversal/Search**: Deep searches to uncover connections.
- **Pathfinding/Connectivity:** Determine shortest or possible paths between nodes.
- **Graph Algorithms**: Support for PageRank, community detection, etc.
- **Data Import/Export**: Easy data integration with other systems.
- **Visualization**: Tools to visually explore relationships.
- **Real-time Analytics**: Immediate insights from real-time querying.

02 Product Review

Leading Products · Technical Details · Marketing Data

Leading Products

		Deployment	Graph Model	Graph OLTP			Graph OLAP	Scale-Out
				Query Language	Visualization tools	Transaction		
Graph Only Companies	**TigerGraph**	On-prem / AWS, Azure, GCP	PG	GSQL	Graph Studio	ACID	GSQL, 23 built-in algorithms	Yes
	Neo4J	On-prem / AWS, Azure, GCP	PG	Cypher	Studio	Non-repeatable reads may occur	Pregel API, 48 built-in algorithms (including Graph ML)	Yes
Data Companies	**DataStax** Enterprise Graph	On-prem / AWS, Azure, GCP	PG	Gremlin	Studio	Row-level (Cassandra)	SparkGraphComputer API	Yes
	Databricks GraphX & GraphFrames	On-prem / AWS, Azure, GCP	PG	Motif Finding DSL	-	-	Pregel API, 7 built-in algorithms	Yes
Enterprise Cloud Companies	**Amazon** Neptune	AWS	PG, RDF	Gremlin, SPARQL	Neptune Workbench	ACID	-	Yes
	Microsoft SQL Graph	On-prem / Azure	PG	SQL Extension	Power BI plugin, 3rd party tools	ACID	Python/R scripts via Machine Learning Services	Yes (Read-Only Queries)
	Microsoft Cosmos DB Graph	Azure	PG	Gremlin	Azure Portal, 3rd party tools	-	-	Yes
	Oracle Spatial and Graph	On-prem / OCI AWS, Azure, GCP	PG, RDF	PGQL, SPARQL	Graph Studio	ACID	Green Marl DSL, 50+ built-in algorithms (including Graph ML)	Yes
	IBM Db2 Graph	On-prem / CP4D	PG	Gremlin	Graph UI	ACID	-	Yes

Tian, Yuanyuan. (2023). The World of Graph Databases from An Industry Perspective. ACM SIGMOD Record. 51. 60-67. 10.1145/3582302.3582320.

Product Deep Dive

- Neo4j
- Amazon Neptune

1

2

1. Hello, World: Neo4j, Inc.: https://neo4j.com/blog/hello-world-neo4j-inc/
2. Amazon Neptune: https://gallery.ecr.aws/neptune/

Neo4j: Deployment Modes

Database as a Service

Cloud-native

Self-service deployment

No access to underlying infrastructure and systems

Cloud Managed Services (CMS)

Customizable deployment model and service levels

Operate in own data centers or Virtual Private Cloud (VPC)

Self-hosted

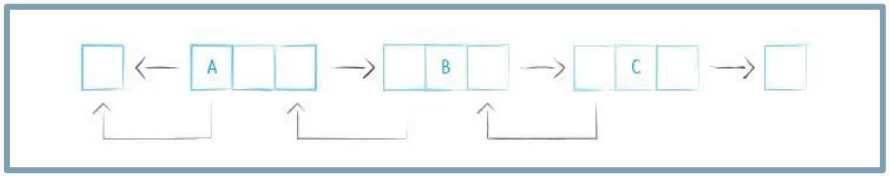
Full control of environment

Run in any cloud

Neo4j Is the Choice of Leading Companies for Graph Databases in the Cloud: https://neo4j.com/press-releases/neo4j-cloud-deployment/

Neo4j: Architecture

- **Native Graph Storage Engine**
  - Index-free adjacency
  - Enhanced performance for pattern-matching & traversals
- **Global Index Types**
  - B-tree indexes
  - Token lookup indexes
  - Full-text indexes



Neo4j Performance Architecture Explained & 6 Tuning Tips: https://www.graphable.ai/blog/neo4j-performance/

Neo4j: Scalability

- **Autonomous Clustering**
  - Automatically allocate copies to the optimal servers
- **Sharding**
  - Divide a single logical database into several smaller databases (shards)
- **Composite Database**
  - Treat distributed graph as a cohesive whole
- **Ops Manager**
  - Easily spin up & manage databases

Neo4j: Security

Role & Schema Based Granular Access Control

- Control users' ability to traverse and read from different parts of the graph.

VPC Isolation

- Database instances and service components are deployed in a separate Virtual Private Cloud (VPC) with dedicated cloud infrastructure.

Data Privacy

- EU General Data Protection Regulation (GDPR) and California Consumer Privacy Act (CCPA) compliant.

Resiliency & Reliability

- 99.95% availability
- Automated Encrypted Backups

Neptune: Graph Data Model

- **Four position (Quad) element**
  - Subject (S)
  - Predicate (P)
  - Object (O)
  - Graph (G)
- **Set of quad statements with shared resources create a graph**
- **Examples**
  - Relationship: Source vertex (S), target vertex (O), edge label (P)
  - Property: Element id (S), property key (P), property value (O)

Neptune: Scalability

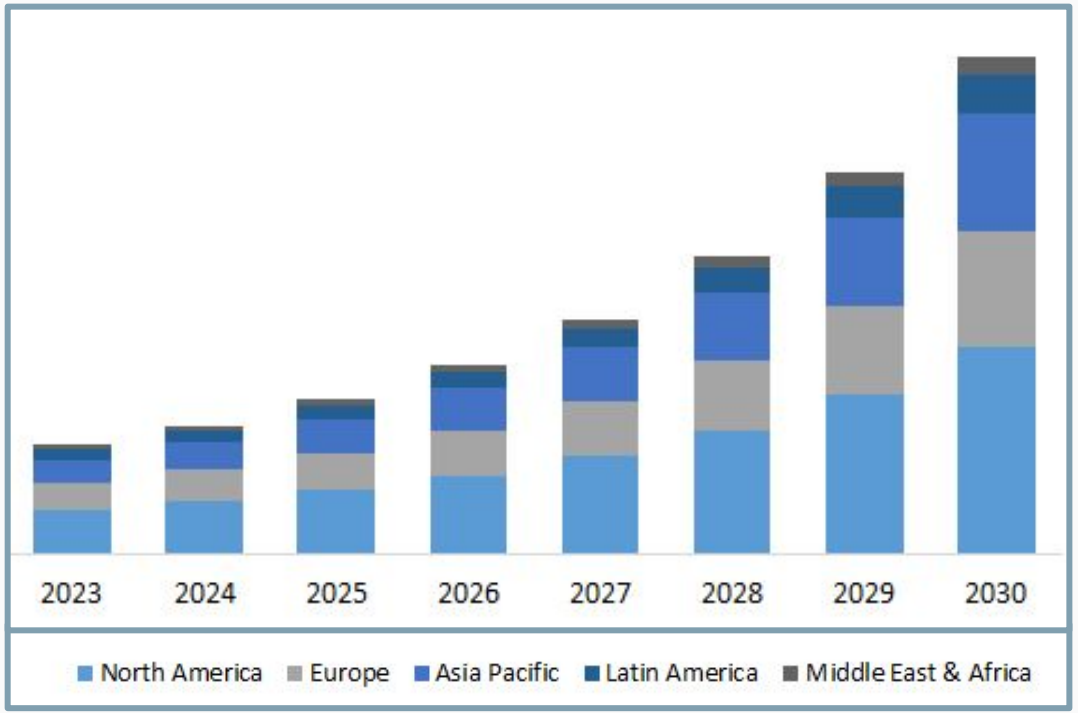
- **Storage Scaling**
  - Automatically grow volume storage (up to 128 TiB)
- **Instance Scaling**
  - Adjust the processing power and memory of your Neptune DB cluster
- **Read Scaling**
  - Up to 15 read replicas per cluster

Neptune: Security

- **Security of the cloud**
  - AWS is responsible for protecting the infrastructure that runs AWS services in the AWS Cloud

- **Security in the cloud**
  - Identity & Access Management
  - Multi Factor Authentication
  - AWS Audit Management
  - Advanced managed security services: Amazon Macie

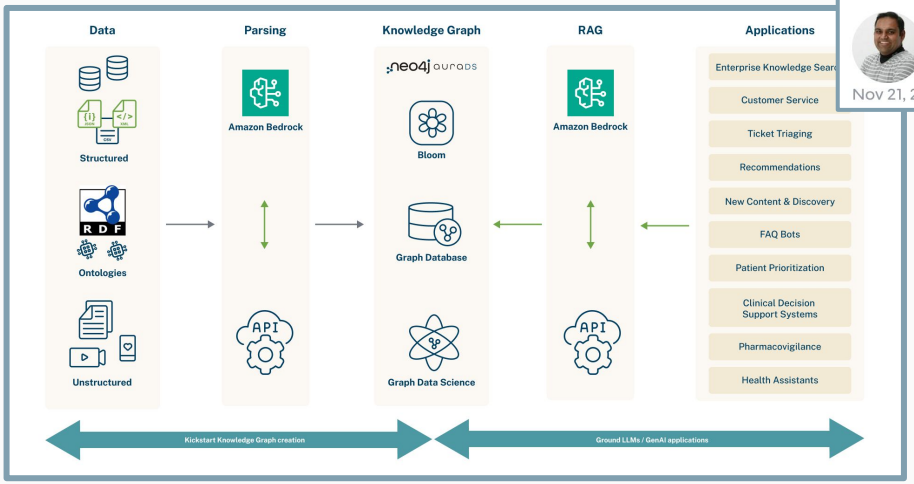
Market Data



- ➔ $2.9 billion in 2023
- ➔ $11.6 billion by 2030.

- Graph Database Market Size Global Report, 2022-20230: https://www.polarismarketresearch.com/industry-analysis/graph-database-market
- Graph Database Market: https://www.marketsandmarkets.com/Market-Reports/graph-database-market-126230231.html

Marketing Strategies



AWS and Neo4j Join Forces to Solve LLM Hallucinations and Evolve GenAI

Sudhir Hasbe, Chief Product Officer, Neo4j

Nov 21, 2023 · 4 mins read

TigerGraph Cloud adds graph analytics, machine learning tools

The updated graph database-as-a-service (DBaaS) will come with visual analytics and machine learning tools, made accessible via the TigerGraph Suite.

- AWS and Neo4j Join Forces to Solve LLM Hallucinations and Evolve GenAI: https://neo4j.com/blog/neo4j-aws-enable-genai/
- TigerGraph Cloud adds graph analytics, machine learning tools: https://www.infoworld.com/article/3680069/tigergraph-cloud-adds-graph-analytics-machine-learning-tools.html

03

Application Research

Sample Applications · Research Problems · Future Trends

Sample Applications

- Fraud Detection [1,2]
- Network & IT Operations
- Identity & Access Management
- Recommendation Engines
- Knowledge Graphs
- Generative AI

Financial Services [3] (UBS, Citibank, Vanguard)

Risk management, compliance reporting, fraud protection, market decision, 360° customer view

Government (US Army, Lockheed Martin, US DHS)

Fight crime, prevent terrorism, improve fiscal responsibility, provide transparency

Life Sciences (Boston Scientific - Medical Supply Chain, Novartis - Drug Discovery)

Pharmaceutical, chemical, biotech data discovery and management

Retail (eBay, Marriott)

Real-time product recommendation, customer experience personalization, supply-chain management

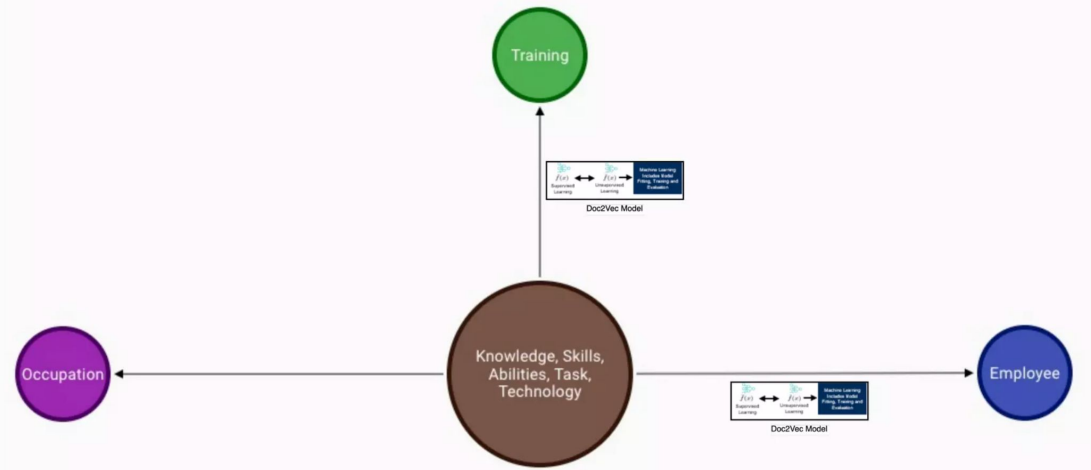
Telecommunication (Cisco, Comcast, Telenor)

Manage complex interdependencies, IT infrastructure, dense network planning

1. Seven Graph Database Use Cases: https://medium.com/technology-hits/7-graph-database-use-cases-that-will-change-your-mind-699e92437523
2. Neo4j Top5 Use Cases: https://go.neo4j.com/rs/710-RRC-335/images/Neo4j_Top5_UseCases_Graph%20Databases.pdf
3. Neo4j Industry Use Cases: https://neo4j.com/use-cases/

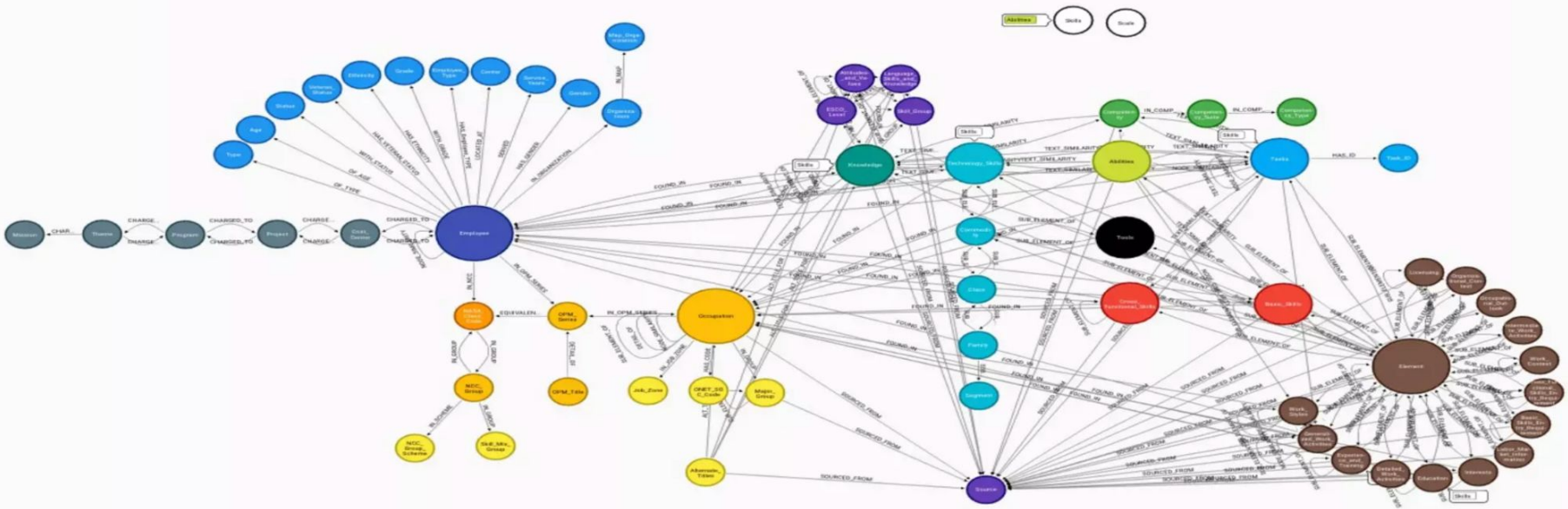
Case Study - NASA (Knowledge Graph)

- Problem: how to connect the right employees to the right roles?
- Goal: break down data silos for efficient information assessment and reasoning
- **Step1: construct graph database**



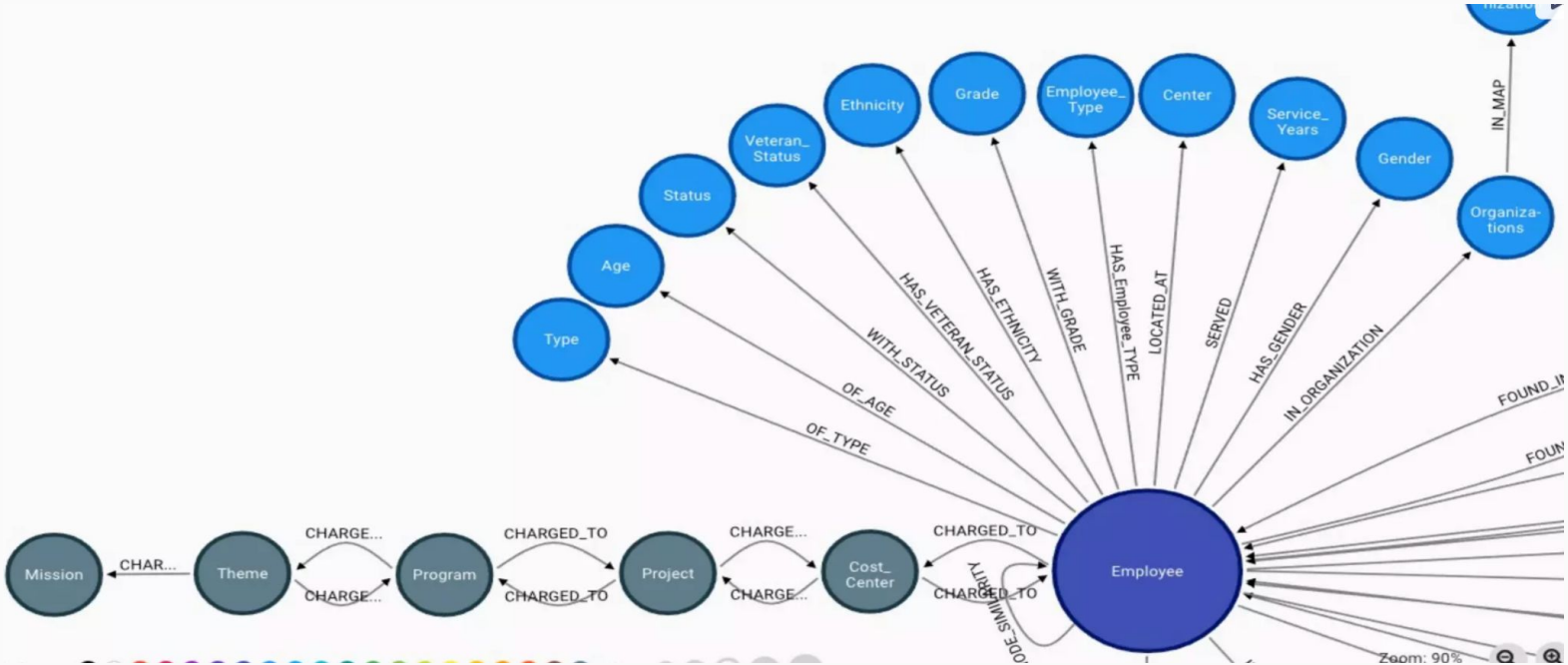
1. Getting to Mars faster with a knowledge graph: https://neo4j.com/case-studies/nasa/
2. Combining knowledge graph and graph algorithms to find hidden skills at NASA: https://neo4j.com/blog/combining-knowledge-graph-graph-algorithms-find-hidden-skills-nasa/

Case Study - NASA (Knowledge Graph)



1. Getting to Mars faster with a knowledge graph: https://neo4j.com/case-studies/nasa/
2. Combining knowledge graph and graph algorithms to find hidden skills at NASA: https://neo4j.com/blog/combining-knowledge-graph-graph-algorithms-find-hidden-skills-nasa/

Case Study - NASA (Knowledge Graph)



1. Getting to Mars faster with a knowledge graph: https://neo4j.com/case-studies/nasa/
2. Combining knowledge graph and graph algorithms to find hidden skills at NASA: https://neo4j.com/blog/combining-knowledge-graph-graph-algorithms-find-hidden-skills-nasa/

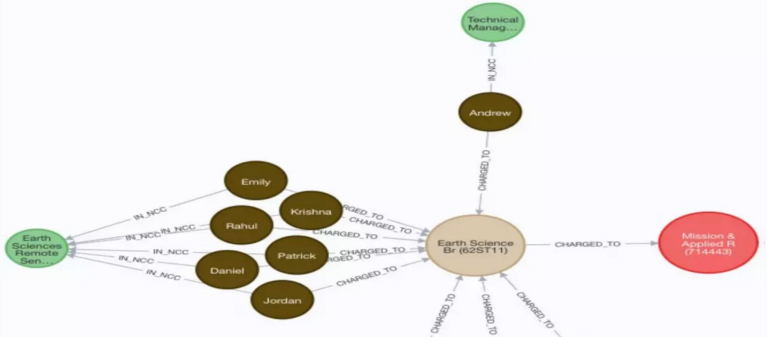
Case Study - NASA (Knowledge Graph)

- Problem: how to connect the right employees to the right roles?
- **Step2: graph data science**
  - Algorithm: node similarity, graph neural network

Use Case 1: Career Change
- **Find similar positions**
- **Identify upscale skill sets**

Use Case 2: Connection Discovery
- **Aggregate project related skills**
- **Analyze project-specific details**

Occupation1	Occupation2	similarity
Management Analysts	Information Technology Project Managers	72.22%
Management Analysts	Document Management Specialists	68.75%
Management Analysts	Sustainability Specialists	68.42%
Management Analysts	Online Merchants	62.50%
Management Analysts	Business Continuity Planners	61.11%
Management Analysts	Health Informatics Specialists	60.00%
Management Analysts	Database Administrators	52.94%
Management Analysts	Information Security Analysts	47.06%
Management Analysts	Web Administrators	47.06%
Management Analysts	Computer Network Support Specialists	44.44%



1. Getting to Mars faster with a knowledge graph: https://neo4j.com/case-studies/nasa/
2. Combining knowledge graph and graph algorithms to find hidden skills at NASA: https://neo4j.com/blog/combining-knowledge-graph-graph-algorithms-find-hidden-skills-nasa/

Case Study - NASA (Knowledge Graph)

"This has saved us at least a year and over $2M in research and development towards our Mission to Mars planning."

— NASA's Lessons Learned Database

1. Getting to Mars faster with a knowledge graph: https://neo4j.com/case-studies/nasa/
2. Combining knowledge graph and graph algorithms to find hidden skills at NASA: https://neo4j.com/blog/combining-knowledge-graph-graph-algorithms-find-hidden-skills-nasa/

Ongoing Research Problems

Database Functionality

- Declarative querying
  - High-level query language
  - Lack of query optimization
- Data partitioning
  - Horizontal scalability
  - Distributed storage
- Data storage architecture
  - Data locality

Big Analytics

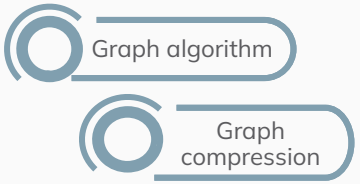
- Graph algorithms
- Graph extraction
  - Extraction from non-graph data
  - Extraction of subgraph
- Graph pattern matching
- Query cost
  - Dynamic temporal graphs
  - Heterogeneous/uncertain graphs
- Parallelisation

Other Challenges

- Need for a benchmark
  - Graph analysis evaluation
  - Scaling and querying
- Graph compression
- Visualization
- Graph stream processing
- Integration & standardization
- Ethical consideration

1. Graph databases: their power and limitations: https://inria.hal.science/hal-01444505/document
2. Future of graph database: https://medium.com/@datumlaviosa/future-of-graph-database-is-it-really-rosy-fdc64e47f8dd

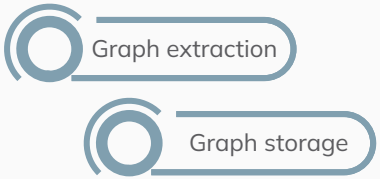
Detecting Anomalous Graphs in Labeled Multi-Graph Databases



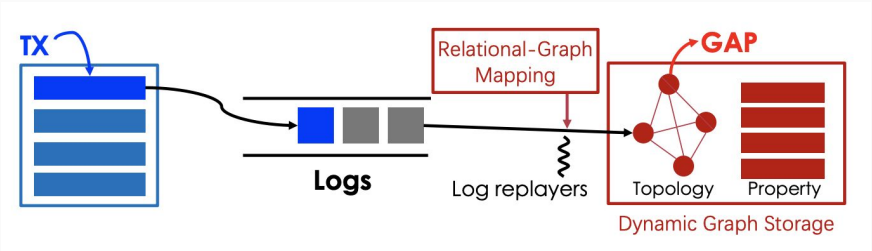
1. Problem formulation
   a. LDM database: node-labeled, directed, multi-graph database
   b. Problem: how to detect anomalous graphs within LDM graph databases
   c. Existing work: anomaly detection designed for unlabeled/unweighted graphs
2. Main contribution
   a. Identify key network motifs (graphical structures) → *search algorithm*
   b. Encode database with identified motifs → *graph encoding schemes*
   c. Flag those graphs that do not compress well under the encoding schema
3. Real-world application
   a. Detecting anomalous employee email graphs with job titles as labels
   b. Control flow graphs with function-calls as labels

1. Detecting anomalous graphs in labeled multi-graph databases: https://dl.acm.org/doi/full/10.1145/3533770

Bridging the Gap between Relational OLTP and Graph-based OLAP



1. Problem formulation
   a. GAP: dynamic graph analytical processing tasks
   b. Problem: balance trade off btw performance degradation and data freshness
2. Main contribution
   a. *GART*: extends hybrid transactional/analytical processing (HTAP) to support GAP
   b. Data model conversion by graph extraction interface → *relational-graph mapping*
   c. Graph storage that supports locality → *dynamic graph storage*

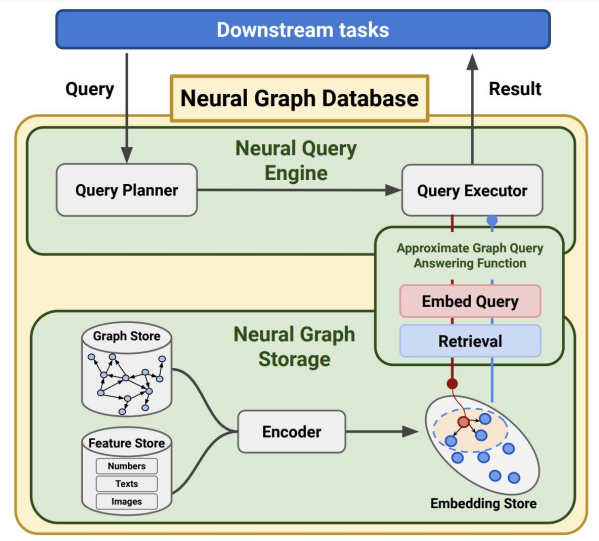


1. Bridging the gap between relational OLTP and graph-based OLAP: https://www.usenix.org/conference/atc23/presentation/shen

Neural Graph Reasoning: Complex Logical Query Answering Meets Graph Databases

Graph storage

Graph + AI



1. Problem formulation
   a. CLQA: solve multi-hop logical reasoning
2. Main contribution
   a. *NGDB*: neural graph databases
   b. *Neural graph storage*: graph store, features store, latent embedding store (encoder)
   c. *Neural query engine*: parameterized approximate graph query answering function
3. Real-world application
   a. Product recommendation, knowledge graph QA

1. Neural graph reasoning: complex logical query answering meets graph databases: https://arxiv.org/pdf/2303.14617.pdf

Future Trends

Graph DB + AI

"The future of graph databases lies at the intersection of AI and advanced data management."

- Intelligent exploration of relations
- Enhanced querying with ML algorithms
- Contextual understanding through NLP
- Predictive analytics for graph data

Continuous Rise of Graph DB

"Graph usage will continue to rise across enterprises."

- Graph can add value in any environment where:
- Data is interconnected and relationships matter
- Data needs to be queried w optimal performance
- Data is evolving and data model isn't prefixed

1. Predictions for the future of graph database: https://www.youtube.com/watch?v=yCOjN8uaIp4&ab_channel=Neo4j
2. Navigating the future: the synergy of AI and graph database: https://dev.to/moiz697/navigating-the-future-the-synergy-of-ai-and-graph-databases-2g9b
3. Why graph DB+AI may be the future of data management: https://www.zdnet.com/article/why-graph-db-ai-may-be-the-future-of-data-management/

Q&A