An Intro to Graphs

Stefan Armbruster

@darthvader42

stefan.armbruster@neotechnology.com

neotechnology
graphs are everywhere

Goal for this talk

- What is Neo4j?
- Why would I use it?
- How do I model data?
- How do I query data?
- What are people using Neo4j for?

Introduction

neotechnology
graphs are everywhere

Relational all the things

VOLUME

COMPLEXITY

neotechnology
graphs are everywhere

The Relational Crossroads

neotechnology
graphs are everywhere

Denormalise

Four NOSQL Categories

arising from the “relational crossroads”

Normalise

neotechnology
graphs are everywhere

HOW TO WRITE A CV

Leverage the NoSQL boom

Let’s talk about graphs

neotechnology
graphs are everywhere

What is a graph?

neotechnology
graphs are everywhere

What is a graph?

neotechnology
graphs are everywhere

Meet Leonhard Euler

- Swiss mathematician
- Inventor of Graph Theory (1736)

Königsberg (Prussia) - 1736

A
B
C
D
1
2
3
4
5
6
7

1
A
4
2
2
1
B
3
C
5
6
5
7
7
D

What are graphs good for?

Complexity

Data Complexity

*complexity = f(**size**, semi-structure, connectedness)*

neotechnology
graphs are everywhere

Size

Name (Symbol)	Value
kilobyte (kB)	10^3
megabyte (MB)	10^6
gigabyte (GB)	10^9
terabyte (TB)	10^{12}
petabyte (PB)	10^{15}
exabyte (EB)	10^{18}
zettabyte (ZB)	10^{21}
yottabyte (YB)	10^{24}

neotechnology
graphs are everywhere

The Real Complexity

complexity = f(size, ***semi-structure***, ***connectedness***)

Semi-Structure

neotechnology
graphs are everywhere

Semi-Structure

USER_ID	FIRST_NAME	LAST_NAME	EMAIL_1	EMAIL_2	FACEBOOK	TWITTER	SKYPE
315	Mark	Needham	mark.needham@neotechnology.com	m.h.needham@gmail.com	NULL	@markhneedham	mk_jnr1984

Email: mark.needham@neotechnology.com
Email: m.h.needham@gmail.com
Twitter: @markhneedham
Skype: mk_jnr1984

neotechnology
graphs are everywhere

The Real Complexity

complexity = f(size, ***semi-structure***, ***connectedness***)

Social Network

neotechnology
graphs are everywhere

Network Impact Analysis

neotechnology
graphs are everywhere

Route Finding

neotechnology
graphs are everywhere

Recommendations

neotechnology
graphs are everywhere

Logistics

neotechnology
graphs are everywhere

Access Control

neotechnology
graphs are everywhere

Fraud Analysis

neotechnology
graphs are everywhere

Neo4j is a Graph Database

When Should I Use Graph Databases??

- Densely-connected, semi-structured domains
  - Lots of join tables? Connectedness
  - Lots of sparse tables? Semi-structure
- Data Model Volatility
- Join Complexity and Performance
- Millions of ‘joins’ per second
- Consistent query times as dataset grows

neotechnology
graphs are everywhere

Graph Modeling

neotechnology
graphs are everywhere

Labeled Property Graph Data Model

Relationships (continued)

Nodes can have more than one relationship

Nodes can be connected by more than one relationship

Self relationships are allowed

Graph Queries

- A language for describing graphs
- Creating nodes, relationships and properties
- Querying data

Querying a Graph

- "Graph local" vs "Graph global"
  - Contextualized "ego-centric" queries
- "Parachute" into graph
  - Start node(s)
    - Found through Index lookups
- Crawl the surrounding graph
  - 2 million+ joins per second
    - No more Index lookups: Index-free adjacency

neotechnology
graphs are everywhere

Queries: Pattern Matching

Pattern

neotechnology
graphs are everywhere

Start Node

Pattern

neotechnology
graphs are everywhere

Match

Pattern

neotechnology
graphs are everywhere

Match

Pattern

neotechnology
graphs are everywhere

Match

Pattern

neotechnology
graphs are everywhere

Non-Match

Pattern

neotechnology
graphs are everywhere

Non-Match

Pattern

neotechnology
graphs are everywhere

Other models to look at

- Graph Gist
  https://github.com/neo4j-contrib/graphgist/wiki
- Chapter 3 of Graph Databases
- Neo4j Manual
  http://docs.neo4j.org/chunked/milestone/data-modeling-examples.html

neotechnology
graphs are everywhere

Graph Visualisations

- Neo4j Browser
- Neography / Neovigator
- Neoclipse
- Gephi, Prefuse, d3.js, etc….

Neo4j Browser

neotechnology
graphs are everywhere

Neoclipse

neotechnology
graphs are everywhere

Neovigator

neotechnology
graphs are everywhere

Other Visualisation options:

- Linkurio.us

- KeyLines

neotechnology
graphs are everywhere

Technical Overview

- Deployment modes
- Java APIs
- Additional libraries

Embedded

- Host in Java process
- Access to Java APIs

neotechnology
graphs are everywhere

Server

- HTTP/JSON interface
- Server wraps embedded instance

neotechnology
graphs are everywhere

High Availability

- Available in Enterprise edition
- Scale horizontally for availability and read throughput
  - Scale vertically for writes
- Master-Slave replication
  - Every instance is full copy of store
- Master coordinates writes
  - Master is immediately consistent
  - Cluster is eventually consistent

Neo4j Architecture

neotechnology
graphs are everywhere

Other Libraries

- Graph Algorithms
  - Shortest Path
  - Shortest Weighted Path
  - A*
  - Dijkstra
  - Custom cost evaluators
  - Available in the core distribution
- Neo4j Spatial
  - Geospatial data
  - 3rd party library
  - Used in Telco production systems
  - https://github.com/neo4j/spatial

Spring Data Neo4j

- POJO based development
- Dynamically generated repositories
- Polyglot persistence
  - Object state persisted to graph and SQL database
  - Distributed transactions
- Maintained by Neo Technology

Case Studies

neotechnology
graphs are everywhere

Industry: Retail
Use case: Retail & C2C Delivery
San Francisco & London

Background

- *As eBay seeks to expand its global retail presence. Quick & predictable delivery is an important competitive cornerstone*
- *To counter & upstage Amazon Prime, eBay acquired U.K.-based Shutl to form the core of a new delivery service, launching eBay Now (www.ebay.com/now) prior to Christmas 2013*
- *Founded in 2009, Shutl was the U.K. Leader in same-day delivery, with 70% of the market*

Business problem

- *Enable customer-selected delivery inside 90min*
- *Maintain a large network routes covering many carriers and couriers. Calculate multiple routing operations simultaneously, in real time, across all possible routes*
- *Scale to enable a variety of services, including same-day delivery, consumer-to-consumer shipping (www.shutl.it) and more predictable delivery times*

Solution & Benefits

- *Neo4j runs at the heart of the system, calculating all possible routes in real time for every order*
- *The Neo4j-based solution is **thousands of times faster than the prior MySQL solution***
- *Queries require 10-100 times less code, **improving time-to-market & code quality***
- *Neo4j makes it possible to add functionality that was previously not possible, and to easily extend the platform over time*

Industry: Media
Use case: Master Data Management (Television EPG Data)
London, UK

Background

- *Zeebox is a well-established UK startup that offers second screen applications to end-users, advertisers and broadcasters*
- *Founded by true media experts, Zeebox aims to reinvent TV since the advent of … TV.*

Business problem

- *Data complexity was growing exponentially as more broadcasters and more shows were being added*
  - *leading to development time increases for applications - a key strategic disadvantage in a fast-moving industry*
- *Query times on the MySQL based model were starting to explode*
  - *risk of having worse end-user experience. This was "make or break" with respect to Zeebox' offering and market position*

Solution & Benefits

- *Neo4j 2.0 offered a much simpler, natural way to model, implement and query their electronic program guide data*
  - *leading to faster development cycles*
  - *no "wedging" of the model into an artificial relational representation*
- *Future-safe solution: adding more channels/broadcasters/programs does not complicate the model unnecessarily*
- *Query times went from 80 seconds (MySQL) to 42 milliseconds (neo4j 2.0 traversal)*

glassdoor™

Industry: Online Job Search
Use case: Social / Recommendations

Sausalito, CA

Background

- *Online jobs and career community, providing anonymized inside information to job seekers*

Business problem

- *Wanted to leverage known fact that most jobs are found through personal & professional connections*
- *Needed to rely on an existing source of social network data. Facebook was the ideal choice.*
- *End users needed to get instant gratification*
- *Aiming to have the best job search service, in a very competitive market*

Solution & Benefits

- *First-to-market with a product that let users find jobs through their network of Facebook friends*
- *Job recommendations served real-time from Neo4j*
- *Individual Facebook graphs imported real-time into Neo4j*
- *Glassdoor now stores > 50% of the entire Facebook social graph*
- *Neo4j cluster has grown seamlessly, with new instances being brought online as graph size and load have increased*

Industry: Communications
Use case: Network Management
Paris, France

Background

- *Second largest communications company in France*
- *Part of Vivendi Group, partnering with Vodafone*

Business problem

- *Infrastructure maintenance took one full week to plan, because of the need to model network impacts*
- *Needed rapid, automated "what if" analysis to ensure resilience during unplanned network outages*
- *Identify weaknesses in the network to uncover the need for additional redundancy*
- *Network information spread across > 30 systems, with daily changes to network infrastructure*
- *Business needs sometimes changed very rapidly*

Solution & Benefits

- *Flexible network inventory management system, to support modeling, aggregation & troubleshooting*
- *Single source of truth (Neo4j) representing the entire network*
- *Dynamic system loads data from 30+ systems, and allows new applications to access network data*
- *Modeling efforts greatly reduced because of the near 1:1 mapping between the real world and the graph*
- *Flexible schema highly adaptable to changing business requirements*

accenture

Industry: logistics
Use case: parcel routing

Background

- One of the world's largest logistics carriers
- Projected to outgrow capacity of old system
- New parcel routing system
  - Single source of truth for entire network
  - B2C & B2B parcel tracking
  - Real-time routing: up to 5M parcels per day

Business Problem

- 24x7 availability, year round
- Peak loads of 2500+ parcels per second
- Complex and diverse software stack
- Need predictable performance & linear scalability
- Daily changes to logistics network: route from any point, to any point

Solution & Benefits

- ideal domain fit: a logistics network is a graph
- Extreme availability & performance with Neo4j clustering
- Hugely simplified queries, vs. relational for complex routing
- Flexible data model reflects real-world data variance much better than relational
- "Whiteboard friendly" model easy to understand

Learning more

neotechnology
graphs are everywhere

Stack Overflow

Find answers or reach to fellow developers with questions.

Ask Neo4j questions »

http://stackoverflow.com/questions/tagged/neo4j

Neo4j Google Group

Share your experiences and expertise with fellow graphistas.

Join now »

http://groups.google.com/group/neo4j

Free Online Course

http://www.neo4j.org/learn/online_course

GraphAcademy

Learn. Graph. Deploy.

Graph Databases Book

www.graphdatabases.com

Neo4j 2.0 by Michael Hunger

http://info.neotechnology.com/Neo4j20_de.html

Meetups / User Groups

Neo4j meetups are worldwide. Make a connection or start a new group.

Join a Meetup »

http://www.meetup.com/Twin-Cities-Neo4j-and-Graph-DB-Meetup

Any questions?

neotechnology
graphs are everywhere