Unleash the power of your meta-data: graph databases

Graphs are everywhere
cfreder2@nd.edu
@cfreder2
Agenda

• What is a Graph Database?
• How are Graph Databases being used today?
• How are Graph Databases different from Relational Databases?
• Introduction to Graph Database Design
• Hands on with the Neo4j Graph Database
• Preview of Notre Dame’s Neo4j backed reporting and data governance portal
What questions can Graph Databases help answer?

- How do I know you?
- How do I get from Notre Dame to Philly?
- What should I eat for dinner tonight?
- What will changing the meaning of this definition break?
- What reports depend on this database field?
- Which report should I use?
Early Adopters of Graph Technology
Detect Criminal Gangs Using Mobile Phone Data

Source: how to detect criminal gangs using graph theory
First-Party Fraud

3 addresses * 3 phones = 9 false identities
What drug will treat this condition but not interact with another drug?
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
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
Are we ready for graph databases?
Why Neo4j?

- Simple to use
- Free
- Open Source
- Active Community
- Frameworks available for every language (.net, ruby, python, java, jdbc, etc...)

Source: Google trends April 28th, 2014
<table>
<thead>
<tr>
<th>Rank</th>
<th>Last Month</th>
<th>DBMS</th>
<th>Database Model</th>
<th>Score</th>
<th>Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td></td>
<td>Oracle</td>
<td>Relational DBMS</td>
<td>1514.08</td>
<td>+22.28</td>
</tr>
<tr>
<td>2.</td>
<td></td>
<td>MySQL</td>
<td>Relational DBMS</td>
<td>1292.67</td>
<td>+2.45</td>
</tr>
<tr>
<td>3.</td>
<td></td>
<td>Microsoft SQL Server</td>
<td>Relational DBMS</td>
<td>1210.43</td>
<td>+5.15</td>
</tr>
<tr>
<td>4.</td>
<td></td>
<td>PostgreSQL</td>
<td>Relational DBMS</td>
<td>230.23</td>
<td>-4.82</td>
</tr>
<tr>
<td>5.</td>
<td></td>
<td>MongoDB</td>
<td>Document store</td>
<td>214.34</td>
<td>+14.35</td>
</tr>
<tr>
<td>6.</td>
<td></td>
<td>DB2</td>
<td>Relational DBMS</td>
<td>184.58</td>
<td>-2.74</td>
</tr>
<tr>
<td>7.</td>
<td></td>
<td>Microsoft Access</td>
<td>Relational DBMS</td>
<td>142.76</td>
<td>-3.72</td>
</tr>
<tr>
<td>8.</td>
<td></td>
<td>SQLite</td>
<td>Relational DBMS</td>
<td>90.17</td>
<td>-2.80</td>
</tr>
<tr>
<td>9.</td>
<td></td>
<td>Cassandra</td>
<td>Wide column store</td>
<td>78.72</td>
<td>+0.63</td>
</tr>
<tr>
<td>10.</td>
<td></td>
<td>Sybase ASE</td>
<td>Relational DBMS</td>
<td>78.14</td>
<td>-3.42</td>
</tr>
<tr>
<td>11.</td>
<td></td>
<td>Solr</td>
<td>Search engine</td>
<td>62.88</td>
<td>+1.74</td>
</tr>
<tr>
<td>12.</td>
<td></td>
<td>Teradata</td>
<td>Relational DBMS</td>
<td>61.73</td>
<td>-0.90</td>
</tr>
<tr>
<td>13.</td>
<td></td>
<td>Redis</td>
<td>Key-value store</td>
<td>58.46</td>
<td>+5.01</td>
</tr>
<tr>
<td>14.</td>
<td></td>
<td>FileMaker</td>
<td>Relational DBMS</td>
<td>54.38</td>
<td>+1.47</td>
</tr>
<tr>
<td>15.</td>
<td></td>
<td>Informix</td>
<td>Relational DBMS</td>
<td>36.70</td>
<td>-0.49</td>
</tr>
<tr>
<td>16.</td>
<td></td>
<td>HBase</td>
<td>Wide column store</td>
<td>36.60</td>
<td>+1.46</td>
</tr>
<tr>
<td>17.</td>
<td></td>
<td>Hive</td>
<td>Relational DBMS</td>
<td>31.02</td>
<td>+0.81</td>
</tr>
<tr>
<td>18.</td>
<td></td>
<td>Memcached</td>
<td>Key-value store</td>
<td>30.99</td>
<td>-1.92</td>
</tr>
<tr>
<td>19.</td>
<td></td>
<td>Elasticsearch</td>
<td>Search engine</td>
<td>29.80</td>
<td>+3.63</td>
</tr>
<tr>
<td>20.</td>
<td></td>
<td>CouchDB</td>
<td>Document store</td>
<td>22.55</td>
<td>-0.31</td>
</tr>
<tr>
<td>21.</td>
<td></td>
<td>Splunk</td>
<td>Search engine</td>
<td>22.51</td>
<td>+0.08</td>
</tr>
<tr>
<td>22.</td>
<td></td>
<td>Neo4j</td>
<td>Graph DBMS</td>
<td>20.56</td>
<td>+1.79</td>
</tr>
<tr>
<td>23.</td>
<td></td>
<td>SAP HANA</td>
<td>Relational DBMS</td>
<td>18.02</td>
<td>+2.10</td>
</tr>
<tr>
<td>24.</td>
<td></td>
<td>Firebird</td>
<td>Relational DBMS</td>
<td>15.68</td>
<td>-0.98</td>
</tr>
<tr>
<td>25.</td>
<td></td>
<td>Netezza</td>
<td>Relational DBMS</td>
<td>15.56</td>
<td>-0.29</td>
</tr>
</tbody>
</table>

Date: April 28th, 2014
100% SPECULATION

Adoption

Relational DB
Document DB
Graph DB

Ignore Ignore Ignore "No" "No" I said "No" dammit "Oh No" "Oh *#!"
Graph DB are classified as NoSQL
How relational databases store records

Customers

Accounts

© All Rights Reserved 2013 | Neo Technology, Inc.
Join Tables: **Customer_Accounts**
Join Tables: Customer_Accounts
In a graph
This is called a Property Graph
Welcome to Hogwarts University
Graph DB 101

A graph database can store any kind of data using a 3 simple concepts:

1. Nodes - graph data records
2. Relationships (or edges) - connect nodes
3. Properties - named data values
A Graph Database

Records in a graph are called Nodes

The simplest graph has just a single node with some named values called Properties. Let's draw a graph that describes Hogwarts University.

Start by creating a Node for our most important Hero: Harry

1. Add the name Harry

2. Let's add his birth year as well.

- Nodes are the name for data records in a graph
- Data is stored as Properties
- Properties are simple name/value pairs (e.g.)
  a. name: Harry
  b. birth_year: 1980
More Nodes

Schema-free, nodes can have a mix of common and unique properties.

Like any database, storing data in Neo4j can be as simple as adding more records. We'll add a few more nodes:

1. Hogwarts is a school, built in 10 or 11 AD.
2. Defense Against the Dark Arts is a skill actively taught at Hogwarts
3. Dumbledore’s Army is a student organization

Note

• Similar nodes can have different properties
• Properties can be strings, numbers, booleans, or arrays.
• Neo4j can store billions of nodes
Organize Nodes into groups

Nodes can be grouped together by applying a Label to each member. In my family graph, we'll label each node that represents a Person:

1. Apply the label "Person" to the node we created for Harry and Dubmledore and color them blue.
2. Apply the label "Organization" to the node we created for D.A. and color it teal.
3. Apply the label "Skill" to the node we created for Defense Against the Dark Arts and color it purple.
4. Apply the label "School" to the node we created for Hogwarts and color it orange.

Notes:

- A node can have zero or more labels.
- Labels do not have any properties.
Add Relationships

Connect Nodes in the graph

The real power of Graphs is in connected data. To associate any two nodes, add a Relationship which describes how the records are related.

In my family graph, we simply say who PARENTS whom:

1. Chris is PARENT_OF Sisaye and Dinkeneh
2. Jane is PARENT_OF Chris
3. Chris HAS_PET Petra

- Relationships always have direction
- Relationships always have a type
- Relationships form patterns of data
Relationship Properties

Store information shared by two nodes

Relationships are data records that can also contain properties. Looking more closely at the relationships, note that:

- Dumbledore was appointed on 01/01/1980 and has no end date for his appointment (yet).
- Harry has reached skill level 3 in defense against the dark arts.
Objectives:

1. Describe the relationship between reports and terms defined in data governance.
2. Describe the relationship between security and reports.
3. Describe the management hierarchy at Hogwarts.
Hands on with Neo4j

1) Download Neo4j ([www.neo4j.org](http://www.neo4j.org))
2) Start the Neo4j Server
3) Download the Hogwarts_bi.cql file from: ([https://github.com/cfreder2/hedwGraphDB](https://github.com/cfreder2/hedwGraphDB))
4) Follow Along at home.
Notre Dame’s BI and Data Dictionary Graph.
Graph Data Model (at the moment)
Active Student

An individual who has been confirmed by an admitting office (or other admitting authority), as recorded by the University Registrar, is considered an active student until he or she:
- Graduates (if degree-seeking)
- Completes the academic term (if non degree-seeking)
- Withdraws or is dismissed by the University
- Fails to enroll for a spring or fall academic term (unless granted a leave of absence by a Dean)

Active Student

Created By: Registrar

This report counts active students based on whether the student is degree seeking and student level. You can explore this report by semester dating back to 2004.

Mendoza College Enrollment

Created By: Registrar

This report counts the number of active students enrolled by Mendoza College of Business. You can explore this report by academic term dating back to 2008.

Undergraduate Enrollment

Created By: Registrar

This report counts active student enrollment for the university over time. You can explore this report by academic year, student level, college, and gender.
Data Driven Decision Making

Report Gallery

**Active Student**
Created By: Registrar
This report counts active students based on whether the student is degree seeking and student level. You can explore this report by semester dating back to 2004.

**Mendoza College Enrollment**
Created By: Registrar
This report counts the number of active students enrolled by Mendoza College of Business. You can explore this report by academic term dating back to 2008.

**Undergraduate Enrollment**
Created By: Registrar
This report counts active student enrollment for the university over time. You can explore this report by academic year, student level, college, and gender.

**Talent Review Dashboard**
Created By: Office of Human Resources
The talent review dashboard provides turnover and affirmative action data for the organizations you are responsible for.
Glossary

Active Dean
An Active Faculty member who holds a current appointment with a Secondary Faculty Appointment Type of Dean, Associate Dean or Assistant Dean, as defined in Article II of the Academic Articles.

Active Department Chair
An Active Faculty member who holds a current appointment with a Secondary Faculty Appointment Type of Chair, Associate Chair or Assistant Chair, as defined in Article II of the Academic Articles.

Active Faculty
An employee of the University who holds an active faculty contract, as defined in Article III of the Academic Articles. Is an active faculty member unless that individual only holds appointments as ROTC faculty.

Credit-Bearing Student
An active student who is enrolled for a given academic term is credit-bearing if he or she is: • Registered in any course for academic credit; or • Registered for a thesis or dissertation course

Dean Type
Indicates the Secondary Faculty Appointment Type of the dean appointment held by an Active Dean.

Faculty Appointment Begin Date
Specifies the date on which a particular Faculty Appointment becomes active.

Non-Returning Student
A non-returning student is a degree-seeking student who either: ● Was an Active Student during the prior academic term and is no longer active during the current academic term and did not graduate; or ● Is a New First Time, Transfer or Returning student during the current academic term who did not complete the academic term as an Active Student.
Active Faculty

An employee of the University who holds an active faculty contract, as defined in Article III of the Academic Articles, is an active faculty member unless that individual only holds appointments as ROTC faculty.

General Information

Data Availability
Data is available for faculty who held appointments from January 2011 to present.

Possible Values
Not Applicable

Notes
Emeritus faculty do not hold current contracts and, therefore, are excluded from the definition of Active Faculty, regardless of any teaching or research responsibilities they may perform. Active Faculty status does not consider leave of absence or compensation status. Individuals providing services in the London, Rome and Dublin programs who do not have an active faculty contract with the University of Notre Dame du Lac are not considered Active Faculty.

R.A.C.I. Matrix

R
President;

A
Campus Data Steward;

C
Research; Human Resources; OSP/R;

I
University Relations; Registrar’s Office; Budget Office; Controllers Office;

Related Reports
1. Undergraduate Enrollment
2. Active Student
3. Menloans College Enrollment
4. Talent Review Dashboard

Last Updated: Mon, 14 Apr 2014 06:30 PM
Active Faculty

An employee of the University who holds an active faculty contract, as defined in Article III of the Academic Articles, is an active faculty member unless that individual only holds appointments as ROTC faculty.

General Information

Data Availability

Data is available for faculty who held appointments from January 2011 to present.

Possible Values

Not Applicable

Notes

Emeritus faculty do not hold current contracts and, therefore, are excluded from the definition of Active Faculty.
Back End:
Data Governance and Reporting Meta-data Repository

Front End:
Search based, quick, mobile friendly

RESTFUL web services

Data Integration (ETL)
This stuff is open source. We are developing this now. The software is currently alpha. Beta will be released in September.

Contact me if you want to contribute: cfreder2@nd.edu or twitter: @cfreder2

Mobile Friendly Business Intelligence Portal
https://github.com/ndoit/huginn
Graph based meta-data repository and restful web services.
https://github.com/ndoit/muninn
Additional Resources

• Download free eBook
• View GraphGists to see how different problem domains can be tackled using graph databases
• Technical mumbo jumbo on Graph vs. Relational (tall table approach) the impact of certain query patterns on performance.