Graph Databases and the Future of Large-Scale Knowledge Management

Marko A. Rodriguez
T-5, Center for Nonlinear Studies
Los Alamos National Laboratory
http://markorodriguez.com

April 8, 2009
Abstract

Modern day open source and commercial graph databases can store on the order of 1 billion relationships with some databases reaching the 10 billion mark. These developments are making the graph database practical for applications that require large-scale knowledge structures. Moreover, with the Web of Data standards set forth by the Linked Data community, it is possible to interlink graph databases across the web into a giant global knowledge structure. This talk will discuss graph databases, their underlying data model, their querying mechanisms, and the benefits of the graph data structure for modeling and analysis.
Outline

- The Relational Database vs. the Graph Database
- The World Wide Web vs. the Web of Data
Outline

• The Relational Database vs. the Graph Database

• The World Wide Web vs. the Web of Data
Our Make Believe World

- Marko is a human and Fluffy is a dog.
- Marko and Fluffy are good friends.
# Our World Modeled in a Relational Database

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Type</th>
<th>Legs</th>
<th>Fur</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>Marko</td>
<td>Human</td>
<td>2</td>
<td>false</td>
</tr>
<tr>
<td>0002</td>
<td>Fluffy</td>
<td>Dog</td>
<td>4</td>
<td>true</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID2</th>
<th>ID2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>0002</td>
</tr>
<tr>
<td>0002</td>
<td>0001</td>
</tr>
</tbody>
</table>

**Object_Table**

**Friendship_Table**
Our World Modeled in a Graph Database

Risk Symposium – Santa Fe, New Mexico – April 8, 2009
Extending our Make Believe World

- Marko is a human and Fluffy is a dog.
- Marko and Fluffy are good friends.
- Human and dog are a subclass of mammal.
## Our Extended World Modeled in a Relational Database

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Type</th>
<th>Legs</th>
<th>Fur</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>Marko</td>
<td>Human</td>
<td>2</td>
<td>false</td>
</tr>
<tr>
<td>0002</td>
<td>Fluffy</td>
<td>Dog</td>
<td>4</td>
<td>true</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID2</th>
<th>ID2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>0002</td>
</tr>
<tr>
<td>0002</td>
<td>0001</td>
</tr>
</tbody>
</table>

### Object Table

- **ID**: 0001, 0002
- **Name**: Marko, Fluffy
- **Type**: Human, Dog
- **Legs**: 2, 4
- **Fur**: false, true

### Friendship Table

- **ID1**: 0001, 0002
- **ID2**: 0002, 0001

### Subclass Table

- **Type1**: Human, Dog
- **Type2**: Mammal

---

Risk Symposium – Santa Fe, New Mexico – April 8, 2009
Our Extended World Modeled in a Graph Database

Risk Symposium – Santa Fe, New Mexico – April 8, 2009
Extending our Extended Make Believe World

- Marko is a human and Fluffy is a dog.
- Marko and Fluffy are good friends.
- Human and dog are a subclass of mammal.
- Fluffy peed on the carpet.
Our Extended Extended World Modeled in a Relational Database

<table>
<thead>
<tr>
<th>ID</th>
<th>Name</th>
<th>Type</th>
<th>Legs</th>
<th>Fur</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>Marko</td>
<td>Human</td>
<td>2</td>
<td>false</td>
</tr>
<tr>
<td>0002</td>
<td>Fluffy</td>
<td>Dog</td>
<td>4</td>
<td>true</td>
</tr>
<tr>
<td>0003</td>
<td>My Rug</td>
<td>Carpet</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ID2</th>
<th>ID2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0001</td>
<td>0002</td>
</tr>
<tr>
<td>0002</td>
<td>0001</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type1</th>
<th>Type2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>Mammal</td>
</tr>
<tr>
<td>Dog</td>
<td>Mammal</td>
</tr>
</tbody>
</table>

Friendship_Table

<table>
<thead>
<tr>
<th>ID1</th>
<th>ID2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0002</td>
<td>0003</td>
</tr>
</tbody>
</table>

Subclass_Table

Pee_Table
Our Extended Extended World Modeled in a Graph Database

Risk Symposium – Santa Fe, New Mexico – April 8, 2009
The Graph as the Natural World Model

- The world is inherently (or perceived as) object-oriented.
- The world is filled with objects and relations among them.
- The multi-relational graph is a very natural representation of the world.
The Graph as the Natural Programming Model

- High-level computer languages are object-oriented.
- Nearly no impedance mismatch between the multi-relational graph and the programming object.
- It is easy to go from graph database to in-memory object.

```java
Human marko = new Human();
marko.name = "Marko";
marko.addFriend(fluffy);
marko.setHasFur(false);
marko.setLegs(2);
```
The Relational Database vs. the Graph Database

- A relational database’s (e.g. MySQL, PostgreSQL, Oracle) data model is a collection interlinked tables.

- A graph database’s (e.g. OpenSesame, AllegroGraph, Neo4j) data model is a multi-relational graph.
SQL vs. SPARQL

SELECT OTY.Name FROM Object_Table AS OTX, Object_Table AS OTY, Friendship_Table WHERE OTX.Name = "Marko" AND Friendship_Table.ID = OTY.ID AND Friendship_Table.Friend = OTX.ID;

SELECT ?z WHERE {
  ?x name "Marko"^^xsd:string .
  ?y friend ?x .
  ?y name ?z }

Risk Symposium – Santa Fe, New Mexico – April 8, 2009
Outline

• The Relational Database vs. the Graph Database

• The World Wide Web vs. the Web of Data
Internet Address Spaces

- The Uniform Resource Identifier (URI) is the superclass of the Uniform Resource Locator (URL) and Uniform Resource Name (URN).
The Uniform Resource Locator

- The set of all **URLs** is the address space of all resources that can be located and retrieved on the Web. URLs denote **where** a resource is.

  - **http://markorodriguez.com/index.html**
  - **Domain name server (DNS):** markorodriguez.com → 216.251.43.6
  - **http://** means GET at port 80,
  - **/index.html** means the resource to get at that Internet location.
The Uniform Resource Name

- The set of all **URNs** is the address space of all resources within the urn: namespace.
  - `urn:uuid:bd93def0-8026-11dd-842be54955baa12`
  - `urn:issn:0892-3310`
  - `urn:doi:10.1016/j.knosys.2008.03.030`

- Named resources need not be retrievable through the Web.

- URNs denote **what** a resource is.
The Uniform Resource Identifier

- The **URI** address space is an infinite space for all Internet resources.
  - urn:issn:0892-3310
  - ftp://markorodriguez.com/private/markos_secrets.txt
  - http://www.lanl.gov#fluffy

- Important: URIs can denote **concepts**, **instances**, and **datum**.

![Diagram of a dog with labels lanl:fluffy and lanl:fluffy_legs]
The “Uniform Resource Graph”

• We can denote where something is, what something is, but how do we denote how something relates to something else?

• How can we denote what something means, where meaning is determined by its place within a larger relational structure?
  * URIs are like words. They denote things in the real or imaginary world.
  * Linking URIs is like defining words. Similar to how a dictionary defines words in terms of other words.
The World Wide Web

- The **World Wide Web** is primarily concerned with the Hyper-Text Transfer Protocol (HTTP) and with retrievable resources in the URL address space.

- These retrievable resources are files: HTML documents, images, audio, etc. The “web” is created when HTML documents contain URLs.

```
index.html
Home.html
Research.html
Resume.html
```

```
http://markorodriguez.com/
```

```
index.html
  href
  `http://markorodriguez.com/`

Resume.html
  `href`
  `http://markorodriguez.com/`

Home.html
  `href`
  `http://markorodriguez.com/`

Research.html
  `href`
  `http://markorodriguez.com/`
```
The Web of Data

- The **Web of Data** is primarily concerned with URIs. If the World Wide Web is the web of files, the Web of Data is the web of data. In other words, for the World Wide Web, the level of granularity is the retrievable file. For the Web of Data, it is the information in that file. Moreover, this information is not necessarily contained in a file. There existence is predicated on their URI. Their meaning is predicated on their relationship to other URIs. The web of URIs is the Web of Data.
The Resource Description Framework

- The Resource Description Framework (RDF) is the standard for representing the relationship between URIs and literals (e.g. float, string, date time, etc.). I would have preferred the name “Uniform Resource Graph” (URG).

- Relationships are directed, labeled links between URIs. A subject URI points to an object URI or literal by means of a predicate URI.

```
subject -- predicate -- object
```

```
lanl:marko -- foaf:knows -- lanl:fluffy
```

```
foaf:name

"Marko A. Rodriguez"^^xsd:string

foaf:name

"Fluffy P. Everywhere"^^xsd:string
```
The RDF Data Model and its Serializations

- RDF is a data model. As such, there exists many serializations (encoding formats) of that model.

- RDF/XML is not RDF. It is a serialization of RDF. It is smart to, at all costs, avoid learning RDF/XML as it is an unintuitive standard. Other serializations include: N-TRIPLE, N3, TRIX, TRIG, ...

The Web of Data is a Distributed Database

- The URI address space is distributed.
- URIs can denote datum.
- RDF denotes the relationships URIs.
- The Web of Data’s foundational standard is RDF.
- Therefore, the Web of Data is a distributed database.
The World Wide Web vs. the Web of Data

[Diagram showing two web servers and a graph database, with links and references to HTML and RDF.]
The Current Web of Data
# The Current Web of Data

<table>
<thead>
<tr>
<th>data set</th>
<th>domain</th>
<th>data set</th>
<th>domain</th>
<th>data set</th>
<th>domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>audioscrobbler</td>
<td>music</td>
<td>govtrack</td>
<td>government</td>
<td>pubguide</td>
<td>books</td>
</tr>
<tr>
<td>bbcblatertotp</td>
<td>music</td>
<td>homologene</td>
<td>biology</td>
<td>qdos</td>
<td>social</td>
</tr>
<tr>
<td>bbcplaycountdata</td>
<td>music</td>
<td>ibm</td>
<td>computer</td>
<td>rae2001</td>
<td>computer</td>
</tr>
<tr>
<td>bbcprogrammes</td>
<td>media</td>
<td>ieeee</td>
<td>computer</td>
<td>rdfbookmashup</td>
<td>books</td>
</tr>
<tr>
<td>budapestbme</td>
<td>computer</td>
<td>interpro</td>
<td>biology</td>
<td>rdfohloh</td>
<td>social</td>
</tr>
<tr>
<td>chebi</td>
<td>biology</td>
<td>jamendo</td>
<td>music</td>
<td>resex</td>
<td>computer</td>
</tr>
<tr>
<td>crunchbase</td>
<td>business</td>
<td>laascnrs</td>
<td>computer</td>
<td>riese</td>
<td>government</td>
</tr>
<tr>
<td>dailymed</td>
<td>medical</td>
<td>libris</td>
<td>books</td>
<td>semanticweborg</td>
<td>computer</td>
</tr>
<tr>
<td>dblpberlin</td>
<td>computer</td>
<td>lingvoj</td>
<td>reference</td>
<td>semwebcentral</td>
<td>social</td>
</tr>
<tr>
<td>dblhnannover</td>
<td>computer</td>
<td>linkedct</td>
<td>medical</td>
<td>siocsites</td>
<td>social</td>
</tr>
<tr>
<td>dblprkexplorer</td>
<td>computer</td>
<td>linkedmdb</td>
<td>movie</td>
<td>surgeradio</td>
<td>music</td>
</tr>
<tr>
<td>dbpedia</td>
<td>general</td>
<td>magnatune</td>
<td>music</td>
<td>swconferencecorpus</td>
<td>computer</td>
</tr>
<tr>
<td>doapsepace</td>
<td>social</td>
<td>musicbrainz</td>
<td>music</td>
<td>taxonomy</td>
<td>reference</td>
</tr>
<tr>
<td>drugbank</td>
<td>medical</td>
<td>myspacewrapper</td>
<td>social</td>
<td>umbel</td>
<td>general</td>
</tr>
<tr>
<td>eurecom</td>
<td>computer</td>
<td>opencalais</td>
<td>reference</td>
<td>uniref</td>
<td>biology</td>
</tr>
<tr>
<td>eurostat</td>
<td>government</td>
<td>opencyc</td>
<td>general</td>
<td>unists</td>
<td>biology</td>
</tr>
<tr>
<td>flickrexporter</td>
<td>images</td>
<td>openguides</td>
<td>reference</td>
<td>uscensusdata</td>
<td>government</td>
</tr>
<tr>
<td>flickrwrappr</td>
<td>images</td>
<td>pdb</td>
<td>biology</td>
<td>virtuososponger</td>
<td>reference</td>
</tr>
<tr>
<td>foafprofiles</td>
<td>social</td>
<td>pfam</td>
<td>biology</td>
<td>w3cwordnet</td>
<td>reference</td>
</tr>
<tr>
<td>freebase</td>
<td>general</td>
<td>pisa</td>
<td>computer</td>
<td>wikicompany</td>
<td>business</td>
</tr>
<tr>
<td>geneid</td>
<td>biology</td>
<td>prodom</td>
<td>biology</td>
<td>worldfactbook</td>
<td>government</td>
</tr>
<tr>
<td>geneontology</td>
<td>biology</td>
<td>projectgutenberg</td>
<td>books</td>
<td>yago</td>
<td>general</td>
</tr>
<tr>
<td>geonames</td>
<td>geographic</td>
<td>prosite</td>
<td>biology</td>
<td>. . .</td>
<td></td>
</tr>
</tbody>
</table>

Los Alamos
NATIONAL LABORATORY
EST. 1943

Risk Symposium – Santa Fe, New Mexico – April 8, 2009
Cultural Differences that are Leading to a New World of Large-Scale Knowledge Management

- Relational databases tend to **not** maintain public access points.

- Relational database users tend to **not** publish their schemas.

- Web of Data graph databases maintain public access points called SPARQL end-points.

- Web of Data graph databases tend to reuse and extend public schemas called ontologies.
Conclusion

• Thank you for your time...

★ My homepage: http://markorodriguez.com
★ Neno/Fhat: http://neno.lanl.gov
★ Faith in the Algorithm: http://faithinthealgorithm.net
★ MESUR: http://www.mesur.org