Database Techniques for Linked Data Management

Andreas Harth, Katja Hose, Ralf Schenkel
Tutorial SIGMOD 2012
Introduction to Linked Data (Andreas)

- Motivation
- Linked Data principles
- Relation to Dataspaces
- Linked Data application architectures
- Conclusion
Centralized storage and query processing (Ralf)

- SPARQL Overview
- Rowstore solutions
- Columnstore solutions
- Other solutions and outlook
Distributed query processing (Katja)

- Motivation for virtual integration
- Lookup-based query processing
- Distributed query processing
MOTIVATION
$ curl -H "Accept: text/turtle"
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix api: <tag:graph.facebook.com,2011:/> .
@prefix og: <http://ogp.me/ns#> .
@prefix fb: <http://ogp.me/ns/fb#> .
@prefix : <http://graph.facebook.com/schema/~/> .
@prefix user: <http://graph.facebook.com/schema/user#> .
@prefix page: <http://graph.facebook.com/schema/page#> .

user:id "100002988319400" ;
user:name "Jesser Weaver" ;
user:first_name "Jesser" ;
user:last_name "Weaver" ;
user:username "jesser.weaver" .

$
Schema.org (Google, Yahoo, Bing)

- Goal: embedding structured data into web pages via microformats
- Popular classes
  - Creative works: CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries ...
  - Embedded non-text objects: AudioObject, ImageObject, VideoObject
  - Event
  - Organization
  - Person
  - Place, LocalBusiness, Restaurant ...
  - Product, Offer, AggregateOffer
  - Review, AggregateRating
Google Rich Snippets/Knowledge Graph

Arizona

Arizona is a state of the United States, located in the southwestern region of the country. Arizona is also part of the Western United States and of the Mountain West states. Wikipedia

Capital: Phoenix

Minimum wage: US$ 7.65 per Hour (January 1, 2012)

Secretary of State: Ken Bennett

Governor: Jan Brewer

Attractions: Grand Canyon, Phoenix Zoo, Arizona Snowbowl, Saguaro National Park, More

Senators: John McCain, Jon Kyl

Report a problem

Shopping results for leica camera

Leica D-LUX 5 10.1 MP Digital Camera (Black) $696
Leica M9 Digital Camera - rangefinder - 147 $6,000
Leica V-LUX 3 12.1 MP Digital Camera $892
Leica 10705 M9 Digital Rangefinder $6,502
Leica X1 12.2 MP Digital Camera (Steel) $1,500
Linked Data on the Web

http://lod-cloud.net/
Types of Data in the Linking Open Data Cloud

http://www4.wiwiss.fu-berlin.de/lodcloud/state/ (Sept 2010)
Billion Triple Challenge Dataset

- Part of the annual Semantic Web Challenge (http://challenge.semanticweb.org/)
- 2011 dataset at http://km.aifb.kit.edu/projects/btc-2011/
- 20GB compressed, 200GB uncompressed
- ~2 bn statements, 213.384 distinct classes, 47.681 distinct properties

<table>
<thead>
<tr>
<th>Class URI</th>
<th># documents</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="http://xmlns.com/foaf/0.1/Person">http://xmlns.com/foaf/0.1/Person</a></td>
<td>1633434</td>
</tr>
<tr>
<td><a href="http://xmlns.com/foaf/0.1/Document">http://xmlns.com/foaf/0.1/Document</a></td>
<td>814800</td>
</tr>
<tr>
<td><a href="http://rdf.freebase.com/ns/common.topic">http://rdf.freebase.com/ns/common.topic</a></td>
<td>572382</td>
</tr>
<tr>
<td><a href="http://www.w3.org/2002/07/owl#Thing">http://www.w3.org/2002/07/owl#Thing</a></td>
<td>468387</td>
</tr>
<tr>
<td><a href="http://purl.org/ontology/mo/MusicArtist">http://purl.org/ontology/mo/MusicArtist</a></td>
<td>346728</td>
</tr>
</tbody>
</table>
WebDataCommons Dataset

- Based on CommonCrawl corpus (http://commoncrawl.org/)
- Parse structured data from HTML pages
  - RDFa
  - HTML Microdata
  - Microformats: hCard, hListing, hCalendar, Geo, hResume, hReview, hRecipe, Species, xfn

<table>
<thead>
<tr>
<th>Crawl date</th>
<th>2009/09-2010/11</th>
<th>2012/02</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total URIs</td>
<td>2.8 bn</td>
<td>1.7 bn</td>
</tr>
<tr>
<td>HTML pages</td>
<td>2.5 bn (28.9TB)</td>
<td>1.5 bn (20.9TB)</td>
</tr>
<tr>
<td>URIs with structured data</td>
<td>148 m</td>
<td>189 m</td>
</tr>
<tr>
<td>Domains with structured data</td>
<td>19 m</td>
<td>65 m</td>
</tr>
<tr>
<td>Resulting statements</td>
<td>5.2 bn</td>
<td>3.3 bn</td>
</tr>
</tbody>
</table>
Scenario Overview

- Semantic Technologies facilitate access to data

1. Query

2. Answer

- Q: data about Berlin?
- Q: famous people that died in Berlin?
- Q: data about Hegel?
- Q: Hegel’s publications?
- Q: data about Marlene Dietrich?
- Q: Dietrich’s songs?
DBpedia

- Linked Data version of Wikipedia
- Scripts that extract data (text, links, infoboxes) from Wikipedia
- Published as Linked Data
- Interlinking hub in the Linked Data web

- Berlin
  - http://dbpedia.org/resource/Berlin

- Hegel
  - http://dbpedia.org/resource/Georg_Wilhelm_Friedrich_Hegel

- Marlene Dietrich
  - http://dbpedia.org/resource/Marlene_Dietrich
BBC Music

- Data about BBC (radio) programmes, artists, songs...
- Combination of BBC-internal data (playlists), MusicBrainz (artists, albums), Wikipedia (artists)
- Underpinning the BBC Music website
- Data published according to Linked Data principles

Marlene Dietrich

- http://www.bbc.co.uk/music/artists/191cba6a-b83f-49ca-883c-02b20c7a9dd5.rdf#artist
Virtual International Authority File (VIAF)

- Joint project of national libraries and related organisations
  - 21 institutions, among them the Library of Congress, Deutsche Nationalbibliothek, Bibliothèque nationale de France
- Provide access to “authority files”
- Matching and interlinking collections from participating institutions

- Hegel
  - http://viaf.org/viaf/89774942/
- Marlene Dietrich
  - http://viaf.org/viaf/97773925/
LINKED DATA PRINCIPLES
Semantic Technologies

- Semantic Web technologies, standardised by the W3C, are mature:
  - RDF recommendation in 1999, update in 2004
  - RDFa (RDF in HTML) note in 2008
  - RDFS recommendation in 2004
  - SPARQL recommendation in 2008
  - OWL recommendation in 2004, update in 2009
  - RIF Core recommendation in 2010
- Linked Data is a subset of the Semantic Web stack, including web architecture:
  - IRI (IETF RFC 3987, 2005)
  - HTTP (IETF RFC 2616, 1999)
Linked Data Principles

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
4. Include links to other URIs, so that they can discover more things.

http://www.w3.org/DesignIssues/LinkedData
1. Use URIs as Names for Things

- Use a unique identifier to denote things
- URIs are defined in RFC 2396

- Hegel, Georg Wilhelm Friedrich
  - http://dbpedia.org/resource/Georg_Wilhelm_Friedrich_Hegel
  - http://viaf.org/viaf/89774942/
  - …

- Hegel, Georg Wilhelm Friedrich: Gesammelte Werke / Vorlesungen über die Logik
  - urn:isbn:978-3-7873-1964-0
Names for Things

“Now! That should clear up a few things around here!”
2. Use HTTP URIs

- Enables “lookup” of URIs
- Via Hypertext Transfer Protocol (HTTP)
- Piggy-backs on hierarchical Domain Name System to guarantee uniqueness of identifiers
- Uses established HTTP infrastructure
- Connects logical level (thing) with physical level (source)
- Important: distinction between “thing URI” and “source URI” („other resource“ vs. „information resource“)
Information Resources vs. Other Resources

Marlene Dietrich, the person

File containing data about Marlene Dietrich

Name?
Creator?
Birth date?
Last change date?
License?
Copyright?

...
Correspondence between thing-URI and source-URI („hash URIs“)

User Agent

http://www.bbc.co.uk/music/artists/191cba6a-b83f-49ca-883c-02b20c7a9dd5.rdf#artist

Web Server

http://www.bbc.co.uk/music/artists/191cba6a-b83f-49ca-883c-02b20c7a9dd5.rdf
Hypertext Transfer Protocol (HTTP)

$ curl -H "Accept: application/rdf+xml" -v
   http://viaf.org/viaf/97773925/

> GET /viaf/97773925/ HTTP/1.1
> User-Agent: curl/7.25.0
> Host: viaf.org
> Accept: application/rdf+xml

< HTTP/1.1 200 OK
< Date: Mon, 28 Mar 2011 17:16:30 GMT
< Content-Location: rdf.xml
< Content-Type: application/rdf+xml; qs=0.9
< Connection: close
Correspondence between thing-URI and source-URI („slash URIs“)

```
User Agent

HTTP GET
303 HTTP GET
RDF

Web Server

http://dbpedia.org/resource/Marlene_Dietrich

http://dbpedia.org/data/Marlene_Dietrich

http://dbpedia.org/page/Marlene_Dietrich
```
3. Provide Useful Information

- When somebody looks up a URI, return data using the standards (RDF*, SPARQL)
- Resource Description Framework, a format for encoding graph-structured data (with URIs to identify nodes/vertices and links/edges)
Resource Description Framework

- Directed, labeled graph
- triple(subject, predicate, object)
  - subject: URI (or blank node)
  - predicate: URI
  - object: URI (or blank node) or RDF literal (string, integer, date…)

- RDF/XML is the most widely deployed serialisation
- Other serialisations possible (N-Triples, Turtle, Notation3…)

- Quadruplets (or quads) used as internal representation when integrating data
- quad(subject, predicate, object, context)
  - context: URI (used to store origin of triple)
Merging Data with RDF


http://example.org/fam#child

http://example.org/smith#brian

http://example.org/smith#carol
4. Link to Other URIs

- Enable people (and machines) to jump from server to server
- External links vs. internal links (for any predicate)
- Special owl:sameAs links to denote equivalence of identifiers (useful for data merging)
Equivalences via owl:sameAs

http://viaf.org/viaf/89774942
- http://dbpedia.org/resource/Georg_Wilhelm_Friedrich_Hegel
- http://www.idref.fr/026917467/id
- http://libris.kb.se/resource/auth/190350
- http://d-nb.info/gnd/118547739
http://www.bbc.co.uk/music/artists/191cba6a-b83f-49ca-883c-02b20c7a9dd5#artist
- http://dbpedia.org/resource/Marlene_Dietrich
http://viaf.org/viaf/97773925
- http://dbpedia.org/resource/Marlene_Dietrich
- http://d-nb.info/gnd/118525565
- http://libris.kb.se/resource/auth/238817
- http://www.idref.fr/027561844/id
http://dbpedia.org/resource/Berlin
- http://mpii.de/yago/resource/Berlin
- http://data.nytimes.com/N50987186835223032381 - Berlin (Germany)
- http://www4.wiwiss.fu-berlin.de/flickrwrappr/photos/Berlin
- http://data.nytimes.com/16057429728088573361 - Gaspe Peninsula (Quebec) (?)
Benefits of Linked Data

- Explicit, simple data representation
  - Common data representation (Resource Description Framework, RDF) hides underlying technologies and systems

- Distributed System
  - Decentralised distributed ownership and control facilitates adoption and scalability

- Cross-referencing
  - Allows for linking and referencing of existing data, via reuse of URIs

- Loose coupling with common language layer
  - Large scale systems require loose coupling, via HTTP as common access protocol

- Ease of publishing and consumption
  - Simple and easy-to-use systems and technologies to facilitate uptake

- Incremental data integration
  - Start with merged RDF graphs and provide mappings as you go
Challenges (I)

- Ramp-up cost for data conversion
  - May be alleviated by semi-automatic mappings and adequate tool support for manual conversion

- Integrated data may be messy at first
  - But can be refined as need arises

- Distributed creation and loose coordination may result in inconsistencies
  - Can be detected, diagnosed, and fixed with appropriate tools
The Pedantic Web Group

- Get the community to contact publishers about errors/issues as they arise
- Get involved: http://pedantic-web.org/
- 137 members!

Acknowledgements to: Aidan Hogan, Alex Passant, Me, Antoine Zimmermann, Axel Polleres, Michael Hausenblas, Richard Cyganiak, Stéphane Corlosquet
Challenges (II)

- Often very much oriented towards individuals
- Little possibilities for expressing schema knowledge
- Different data sources have different ways of representing the same facts

- Ontology languages (RDFS, OWL) solve that drawback
- RDFS and OWL are layered on top of RDF
Dataspaces

- Abstraction for Data Management to overcome data integration problems
Dataspace Architecture Components

- Catalog and browse
  - Collection of data sources (schema, rate of change, accuracy…)

- Search and query
  - Query everything
  - Structured queries
  - Metadata queries
  - Monitoring

- Local store and index
  - Store associations between objects, increase availability,…

- Discovery
  - Locate new databases

- Source extension
  - Add query functionality,…
Linked Data vs. Dataspaces

- Method for decentralised data publishing and interlinking
- Ecosystem (incl. people)
- m:n mappings
- Many small sources
- Decentralised interlinking
- No central catalog

- Comprehensive architecture for data integration
- Platform
- 1:m mappings
- Few large sources
- Links in the local index
- Central catalog
LINKED DATA APPLICATION ARCHITECTURES
Architecture Styles

Warehousing/
Crawl-Index-Serve

Virtual Integration/
Distributed Querying

0. Crawl-Index

1. Query
2. Answer

? !

? !
Basic Application: Entity Browsing

**Warehousing/ Crawl-Index-Serve**

Google, SWSE, Falcons, Sindice, Watson, FactForge...

**Virtual Integration/ Distributed Querying**

Tabulator, Disco, Zitgist...
SUMMARY
Summary

- The Linked Data Web is a large, decentralised, complex system built on simple principles
  - identify resource via HTTP URIs
  - provide RDF that links to other URIs upon lookup
- Current trend around Linked Data allows for a re-think of components in Semantic Web Layer Cake
- Data publishers and consumers coordinate little
- Web of Data grows rapidly and covers a large variety of domains
- Algorithms operating over a common access protocol and data model
- Ontology languages provide integration and mapping between disparate sources
- First commercial applications emerging
Attribution

- Slides adapted from my SWT-2 lectures and WWW 2010 SILD and INFORMATIK 2011 tutorials
- Images of Berlin, Hegel and Dietrich via Wikipedia
Motivation

- With increased use of computers more and more data is being stored
  - Organisations rely on data for business decisions
  - Data drives policy decisions in government
  - Individuals rely on data from the Web for information and communication

- Data volumes explode
  - More and more data available on the Web is represented in Semantic Web standards
  - Linking Open Data (LOD) initiative

- Semantic Web technologies facilitate the integration of data from multiple sources
- Combining data from multiple sources enables insights
RDF Graph collected via breadth-first expansion from http://danbri.org/foaf.rdf

7683 triples from 25 RDF files
1062 IRIs
154 blank nodes
1160 literals
Information Resource Graph

RDF graph collected via breadth-first expansion from http://danbri.org/foaf.rdf
319 nodes from RDF files
453 edges

average outdegree: 25
http://mmt.me.uk/foaf.rdf has outdegree of 105!
Dataspace Architecture
Semantic Web Components

User Interface & Applications

Query: SPARQL

Data interchange:
  RDF
  XML
  URI/IRI
Linked Data: Minimal Components

1. Query
2. Answer

User Interface & Applications

Query: SPARQL
Data interchange:
RDF  XML  URI/IRI

1. Query
2. Answer
Data Integration System Architecture

Integration

Wrapper 1
Source 1

Wrapper 2
Source 2

Wrapper n
Source n
Linked Data on the Web
Linked Data on the Web

2008-03
Linked Data on the Web

As of July 2009

2009-07