SHARD Triple-Store:
Tools for Web-Scale SemWeb

Kurt Rohloff
krohloff@bbn.com
@avometric

Many thanks to:
Mike Dean, Ian Emmons, Gail Mitchell, Doug Reid, Rick Schantz from BBN
Hanspeter Pfister from Harvard SEAS
Phil Zeyliger from Cloudera
Prakash Manghwani
Semantic Web / Graph Data

• Vision from Tim Berners-Lee at W3C.

• Create a web of data
  – Support use by intelligent agents.
  – Data described using ontologies.
  – Data represented as digraphs.
  – “Web 3.0.”

• Emerging commercially
  – Use by NYTimes, BBC, Pharma, …
  – Numerous startups.
  – Oracle, MySQL have SemWeb support.

• Government use…
Object Graph Example

BBN

City

Massachusetts

Cambridge

Person

elmer

president

locatedIn

US

Country

State

locatedIn

Company

Organization

rdfs:subClassOf

rdfs:subClassOf

name

“BBN Technologies”

“Tad Elmer”

Company

 RDF: type

Organization

rdfs:subClassOf

Country

US

locatedIn

State

locatedIn

City

locatedIn

name
SemWeb Layer Cake

User interface and applications

Trust

Proof

Unifying logic

Ontologies: OWL
Rules: RIF/SWRL

Taxonomies: RDFS

Data interchange: RDF

Syntax: XML

Identifiers: URI
Character set: UNICODE

Querying

Knowledge Storage

Reasoning

Cryptography
W3C Resource Description Framework (RDF)

- RDF graph is made up of individual statements.
- Subject and predicate are Uniform Resource Identifiers (URIs).
- You can also make statements about statements (e.g. timestamp, confidence, etc.)
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns="http://example.org/business-ont#">
  <Company rdf:ID="BBN">
    <name>BBN Technologies</name>
    <headquarters rdf:resource="http://www.state.ma.us/cities#Cambridge"/>
    <president rdf:resource="http://www.bbn.com/management#elmer"/>
  </Company>
</rdf:RDF>
All people who own a car made in Detroit:

```
SELECT ?person
WHERE {
  ?car a :Car .
  ?car :madeIn :Detroit .
}
```
Answering Queries

Kurt owns car0 madeBy Ford madeIn Detroit livesIn Cambridge.

?person owns ?car madeIn Detroit.
Sample of Triple-Stores

- Parliament by BBN (from DAPRA DAML.)
- OWLIM by OntoText (several versions.)
- Allegrograph from Franz.
- MySQL and Oracle Solutions.
- LarKC by DERI Galway.
- Mulgara.
- Hive- and Pig-based experimental triple-stores.
- Etc…
Triple-Store Design Considerations

- Scalable – web-scale?
- High Assurance.
- Cost Effective – commodity hardware?
- Modular inferred data separation.
- Robustness.

- Considerations as endless as applications.
Map-Reduce Triple-Store Proof of Concept

STAND BACK
I'M GOING TO TRY SCIENCE
SHARD Triple-Store Built on Hadoop

Prioritized goals:

• Commodity hardware, ONLY.
• Web scalable.
• Robust.
More Specifically

- Cloud-based triple-store on HDFS.
  - Method calls at client.
  - Processing in cloud.
  - Move results to local machine.

- Massively scalable.
- SPARQL queries.
- Basic inferencing.
Data Persistence Advice from SHARD

- Down to “bare metal” in HDFS for efficiency.
  - No Berkeley DB, no C-stores, …. Nothing.

- Simple data storage as flat files.
  - Lists of (predicate, object) pairs for every subject by line.
  - Ex: Kurt owns car0 livesin Cambridge

- Simple often really is better…
HDFS Graph Storage

Graphs saved as flat-file in HDFS:

Kurt owns car0 livesIn Cambridge
Car0 a Car madeBy Ford madeIn Detroit
Cambridge a City
Detroit a City
Query Processing

• BBN-developed query processor.
  – Starting integration with “standard” interfaces
    • Jena, Sesame.

• SHARD supports “most” of SPARQL.
  – Like most commercial triple-stores.

• Large performance improvements possible with improved query reordering.
Iterative Query Response Construction

Source Data

1st clause results

2nd clause results

2nd clause results
Test Data

• Deployed code on Amazon EC2 cloud.
  – 19 XL nodes.

• 6000 LUBM university dataset.
  – Approximately 800 million edges in graph.

• In general, performed comparably to “industrial” monolithic triple-stores.
SHARD Open-Source Release

• BSD license.
• Check:
  – My webpage
  – Sourceforge (SHARD-3store)
More info?

- Tim Berners-Lee’s seminal SciAmerican article.
- W3C for “recommended” standards.
- Jena and Sesame frameworks.
- SemWebCentral for other open-source.

- Please come up and talk with me for more info!
Thanks!

Questions?

Kurt Rohloff
krohloff@bbn.com
@avometric
Performance Comparison

• Proof o’ Concept: For 6000 universities (approx. 800 million triples):
  Query 1: 404 sec. (approx 0.1 hr.)
  Query 9: 740 sec. (approx 0.2 hr.)
  Query 14: 118 sec. (approx 0.03 hr.)

• Sesame+DAMLDB:
  Query 1: approx 0.1 hr,
  Query 9: approx 1 hr
  Query 14: approx. 1 hr

• Jena+DAMLDB for 550 million triples:
  Query 1: approx 0.001 hr,
  Query 9: approx 1 hr
  Query 14: approx. 5 hr