SemWeb Challenges:
Scalability, Performance, Scalability, Cost-Effectiveness and Scalability

Kurt Rohloff
krohloff@bbn.com
@avometric

Raytheon
BBN Technologies

A History of Innovation

Raytheon BBN Technologies

1950s

Acoustic Design for UN General Assembly Hall

AI Program for Pattern Recognition

1960s

Demonstration of Time Sharing

LOGO Programming Language

ARPANET-First Multi-node Packet Switched Network

1970s

First Person-to-Person Network Email

@ Sign for Email Addresses

Acoustic analysis of JFK Assassination Tapes

Analysis of Nixon Watergate Tapes

First Symmetric Multi-processor

First TCP for UNIX

1980s

First Electronic Mail

Defense Data Network

Natural Language Computer Interface

Intelligent Agents

SimNet

Collaboration Planning Technology

1990s

Secure email for DoD

DARPA Information Assurance

Broadband Wireless Technology

Genetic Algorithm Scheduling Tools

Collaborative Planning for Desert Storm

ATM Switch

40K Word Speech Recognition System

Safekeyper Certificate Management

Certificate Authority Workstation (CAW)

2000s

Call Director Natural Language Routing

DARPA Agent Markup Language

Semantic Web – OWL, SWRL, OWL-S

Asio tools for Net-Centric Data Sharing

Ultra*Log Agent-Based Network Survivability

Boomerang Mobile Shooter Detection System

Quantum Cryptographic Network

Tools and Benefits

Raytheon
BBN Technologies

- Parliament™ RDF Store
- Asio™ Query Federator
- SILK Reasoning Engine
- Deconfliction
- Process model matching
- Semantic similarity
- Clustering & graph analysis

- Application-independent data
- Query across disparate data sources
- Reason formally about policy and process compliance
- Match process models to observed events
- Find “non-obvious relationships” in large volumes of linked data

Challenges

Raytheon
BBN Technologies

SCALABILITY!!!!

Also:

- High-performance query processing.
- Inferencing.
- Support multiple standards-based interface(s).
- Meaningful benchmarks.

Why?

Raytheon
BBN Technologies

- Triple-Store Study:
  - "An Evaluation of Triple-Store Technologies for Large Data Stores", SSWS '07 (Part of OTM)
  - Great help from OntoText, Franz
- Design Goals (not just scalability!):
  - Scalable – avoid monolithic resource limitations.
  - High Assurance – maintain QoS despite major failures.
  - Cost Effective – only commodity hardware.
  - Modular – strong data separation to maintain provenance

SHARD: SemWeb Cloud Concept

Raytheon
BBN Technologies

A Preface

Raytheon
BBN Technologies

SHARD is released open-source.

- BSD license.
- Look at:
  - My webpage (Search for “SHARD krohloff”)
  - Sourceforge (SHARD-3store)
- Use svn to get code:

```
svn co https://shard-3store.svn.sourceforge.net/svnroot/shard-
   3store shard-3store
```

  - Don’t worry - this command is on SourceForge!

SHARD v01 Operation Overview

Raytheon
BBN Technologies

- Method calls at client.
  - Rolled-our-own query engine.
- SPARQL processing via MapReduce jobs.
- Move results to local machine for local drill-down.
  - Hadoop abstraction layer manages partial system failures with storage and computation redundancy

Test Data

Raytheon
BBN Technologies

- Standard LUBM benchmark data
  - Artificial data on students, professors, courses, etc… at universities
- Deployed code on Amazon EC2 cloud
  - 19 XL nodes
- 6000 university dataset
  - Approximately 800 million edges in graph
- In general, performed comparably to "industrial" monolithic triple-stores

General Programming for Scalable Cloud Computing

Raytheon
BBN Technologies

From Experience:

- Inherently multi-threaded.
- Toolsets still young.
  - Not many debugging tools.
- Mental models are different...
  - Learn an algorithm, adapt it to chosen framework.
  - Ex: try to fit problem into PageRank design pattern.
    - (This isn't what we do, but this approach seems common.)

Raytheon
BBN Technologies

Thanks!
Questions?
We're Hiring!!

Kurt Rohloff
krohloff@bbn.com
@avometric