Data Ontology Cache System

Derek Mansen, Marc Weil, Matt Kotsenas, Robbie Gladmon, Tom Rudick
Two Sigma Investments

• Process-driven investment firm
• Manages billions in assets
• Analyzes data to determine investment strategies
Problem

• Large datasets
  - Different formats
  - Different datastores

• Find relationships between data

• Perform queries across datasets
What’s an Ontology?

- Flexibly structured data
  - Graph representation
  - Graph vs Relational vs Hierarchical
- Explore and expose relationships
  - Computer can read, reason and write
Why an Ontology?

Market

Company

Vendor

Dataset

Investment

hasComponent
componentOf

vends
vendedBy

containsCompany

influencesInvestment

investedin

derivedFrom
System Architecture

- Collectors
- Ontology Core
- Query Clients
Collectors

- Flexible API
- Outsources domain knowledge
- Run on demand
Ontology Core

• Stores data cache
• Exposes relationships
• Executes queries
Ontology Core

• Singular data store
• Content agnostic
• Replicated
  ▪ Availability
  ▪ Performance
Query Clients

• Multiple end-points
  ▪ Web interface
  ▪ Command line interface
  ▪ Groovy interface

• Well-defined API
Query Clients

- **API**
  - Uses SPARQL
- **Web**
  - Easy-to-use graphical front-end
- **Command line**
  - Integrates easily into existing tools
- **Reports**
  - Groovy provides scripting support
Demo
Project Status

• System already in use
• Documented known issues
• Stubbed security classes
Difficulties

• Testing
  ▪ Glue code
  ▪ Automation

• Performance
  ▪ Test data generation
  ▪ Tool limitations

• Replication
  ▪ Hard
  ▪ Testing is even harder
Accomplishments

• Solid architecture
• Little rework
• Learned about the Semantic Web
• Effectively handled large scope
Questions?