bigdata®

Managing Scale in Ontological Systems

SYSTAP Company Overview

Overview

- LLC, Small Business, Founded 2006
- 100% Employee Owned, 2 Principals
- 35 Years Combined Experience, 16 Years With Semantic Web Technologies

Customers & Use Cases

- **Intelligence Community**
  - Federation and semantic alignment at scale to facilitate rapid threat detection and analysis
- **Telecommunications**
  - Horizontal data integration across enterprise services
- **Health Care**
  - Data integration and analytics
- **Network Storage**
  - Embedded device monitoring and root cause analysis
- **Collaboration and Knowledge Portals**
  - Bioinformatics, manufacturing, NGOs, etc.
- **OEM Resellers**

Corporate Services & Product Offering

- **Semantic Web Consulting Services**
  - System vision, design, and architecture
  - Information architecture development
  - Ontology development and inference planning
  - Relational data mapping and migration
  - Rapid prototyping
- **Bigdata®, an open-source, horizontally-scaled high-performance RDF database**
  - Dual licensing (GPL, commercial)
  - Infrastructure planning
  - Technology identification and assessment
  - Benchmarking and performance tuning
  - Feature development
  - Training & Support

What is “big data?”

- Big data is a way of thinking about and processing massive data.
  - Petabyte scale
  - Distributed processing
  - Commodity hardware
  - Open source

Different kinds of “big” systems

- Row stores
- Map / reduce
- Main memory graph processing
  - Boutique super computers, Cray XMT, etc.
- Parallel (clustered) databases
  - The Bigdata® platform fits into this category.

Timeliness vs. Completeness

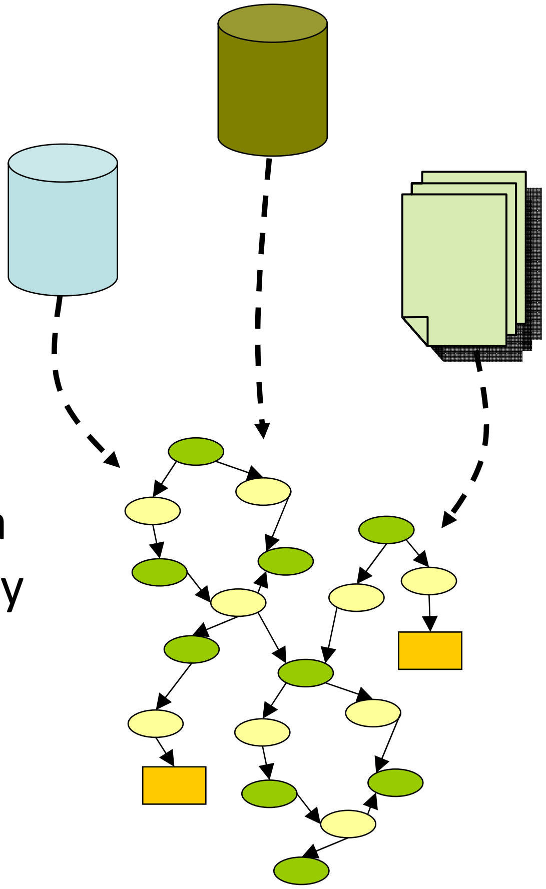
- Rapidly exploit fusion of data sources.
  - Exploitation cycle can be just a few hours.
- High level reasoning over curated information
  - Careful, detailed, and length period of ontology development;
  - In depth reconciliation of data sources and their semantics.
  - Exploitation cycle can be six months to several years.

Expressivity vs. Scale

- Don’t be seduced by expressivity
- Computationally expensive
- High expressivity not easily partitioned
- A little ontology goes a long way
- Avoid constructs that tell you things you probably already know (e.g. domain/range)

The killer "big data" app

- Clouds + "Open" Data = Big Data Integration
- Critical advantages
  - *Fast* integration cycle
  - Open standards
  - Integrate heterogeneous data, linked data, structured data, and data at rest.
  - Opportunistic exploitation of data, including data which can not be integrated quickly enough today to derive its business value.
  - Maintain fine-grained provenance of federated data.



Information Architecture

- Provenance
  - Bigdata® has a dedicated mode for datum level provenance. Fast, inline representation with SPARQL query and only 20% of the foot print on the disk.
- Modeling relationships
  - Provenance model allows dual modeling of relationships as entities.
- Benefits of micro ontologies
  - Separate out system architecture, application architecture, and domain architecture.

CAP Theorem

- Distributed systems can have at most 2 out of 3:
  - Consistency
  - Availability
  - Partition Tolerance
- Bigdata sacrifices *Consistency*
  - Updates are *shard-wise ACID*
  - Application level protocols can provide globally consistent updates

Cloud Architecture

- Hybrid shared nothing / shared disk architecture
  - Compute cluster
    - Spin compute nodes up or down as required
  - plus
  - Managed cloud storage layer
    - S3, openstack, parallel file system, etc

bigdata ®

Flexible
Reliable
Affordable
Web-scale computing.