SEMANTIC TECHNOLOGY & BUSINESS CONFERENCE

10th ANNUAL

Case Study: High-Scale Semantics from Text to Wiki

Mark Wallace

Modus Operandi, Inc.

In Support of Office of Naval Research Contracts:
SINEW - Synchronization Infrastructure for Networked Enterprise Warfare - Contract #: N00014-12-C-0369
KASL III – Knowledge Acquisition Service Library – Contract #: N00014-10-C-0332

PRODUCED BY:

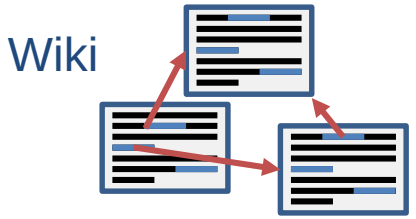
DATAVERSITY

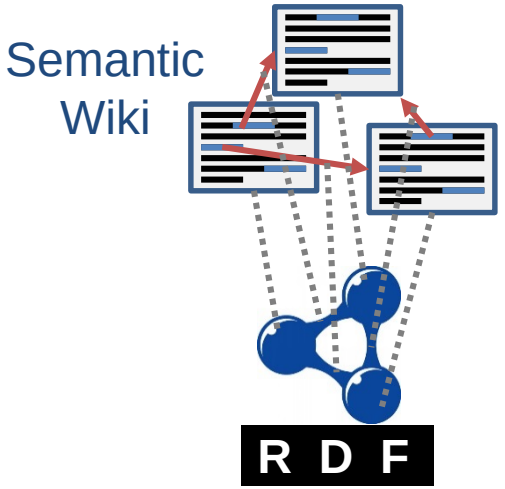
semanticweb.com

#SemTechBiz

The Product to Scale: Semantic Wiki

- Wiki = collaborative web site, edited by non-webmasters
- Semantic Wiki = wiki that maintains a semantic knowledge model under-the-hood
  - Pages become graph nodes
  - Links between pages become edges between nodes
  - Key text on pages become data values on nodes
  - Dynamic page content (via model queries)
- Pages could be generated by *ingesting* documents and performing *entity extraction*





#SemTechBiz

SEMANTIC TECHNOLOGY & BUSINESS CONFERENCE

The Challenge: Scale 10x

AS-IS

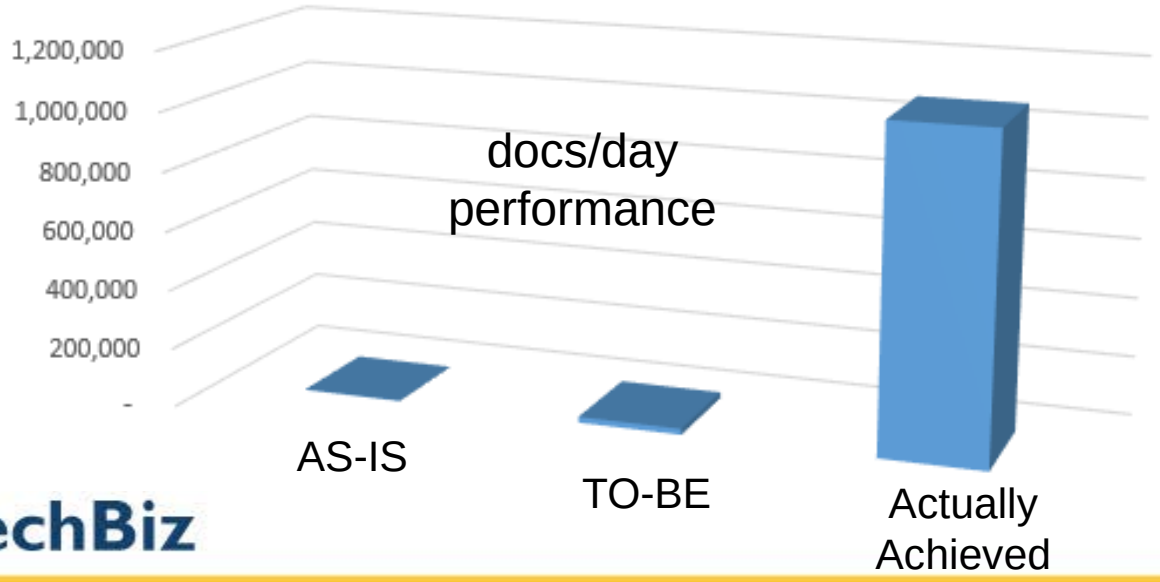
- 2000 docs/day
- Store 0.5 M docs
- 0.1 B Triples

TO-BE

- 20,000 docs/day
- Store 6 M docs
- 2 B Triples

Actually Achieved

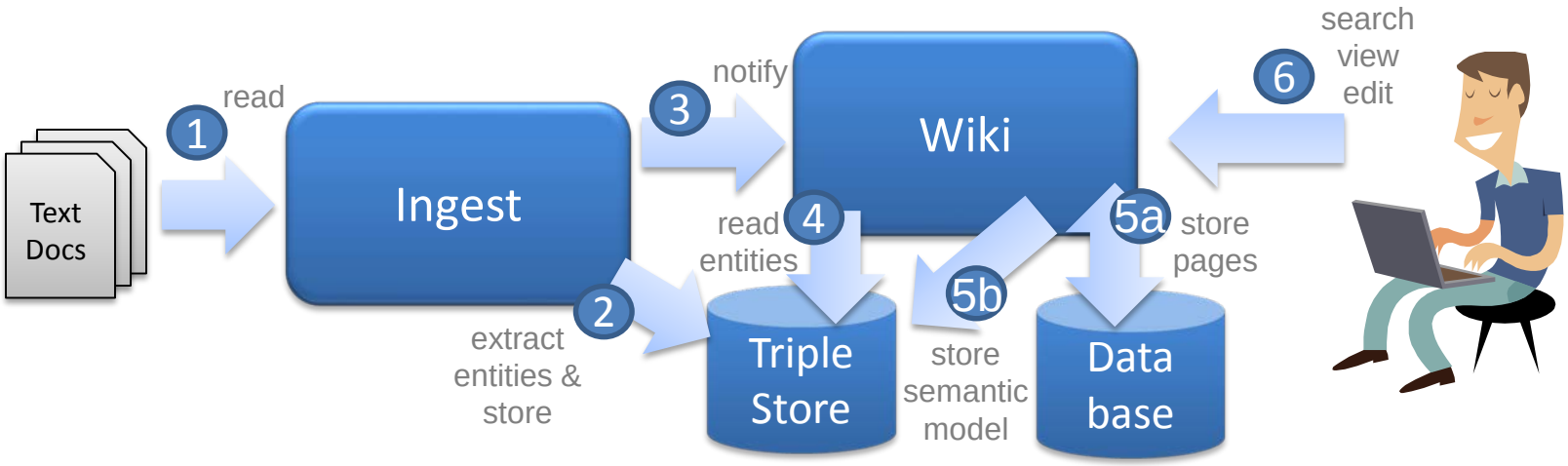
- 45,000 docs/hr
- Store 6 M docs
- 2 B Triples



#SemTechBiz

SEMANTIC TECHNOLOGY & BUSINESS CONFERENCE

Basic Parts of the System



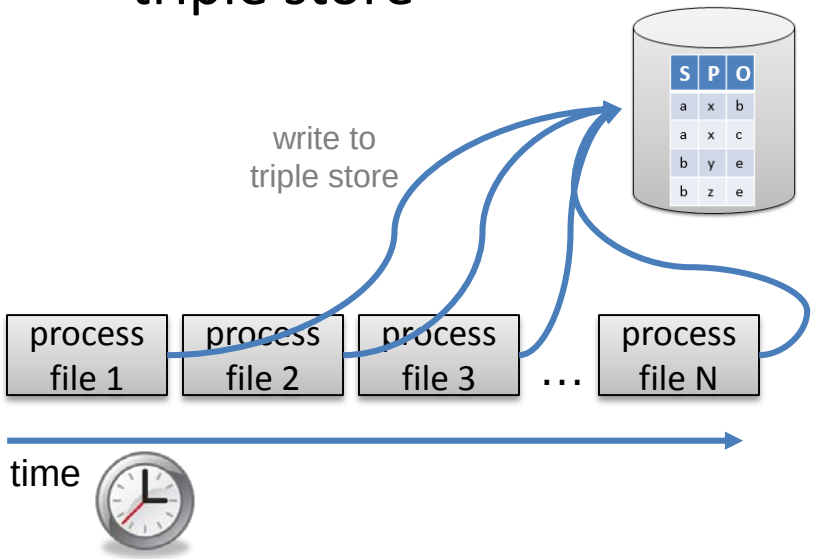
- **Ingest** - ingests documents, extracts entities, notifies wiki
- **Wiki** - auto-creates pages from entities; search, view, edit pages
- **Triple Store (TS)** - stores semantic models
- **Database (DB)** - stores wiki pages

Architecture Changes to Improve Scaling

Hadoop®-ify Ingest Processing

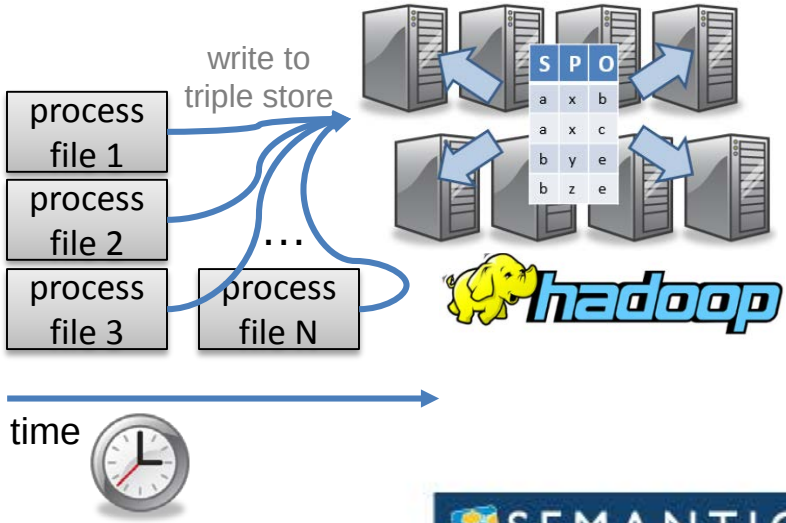
Before

- Serial file processing
- Writing to single node triple store



After

- Parallel file processing
- Writing to cluster-based triple store (Rya)



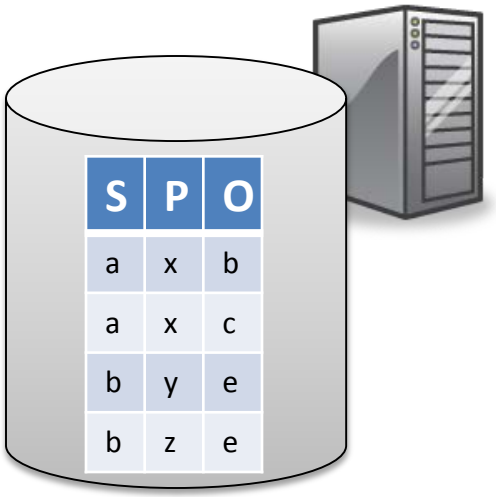
#SemTechBiz

SEMANTIC TECHNOLOGY & BUSINESS CONFERENCE

Move to Scale-out Triple Store

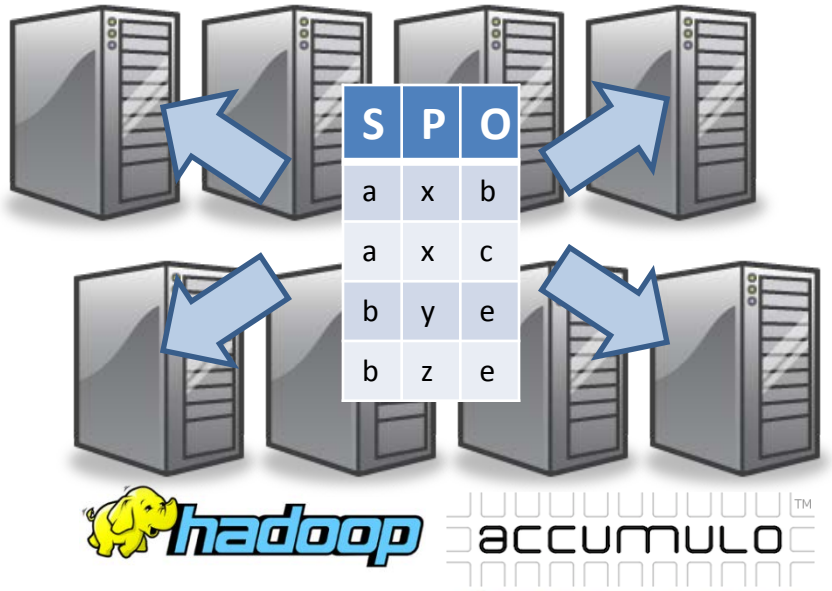
Before

- Single-node triple store



After

- Cluster-based triple store (Rya)



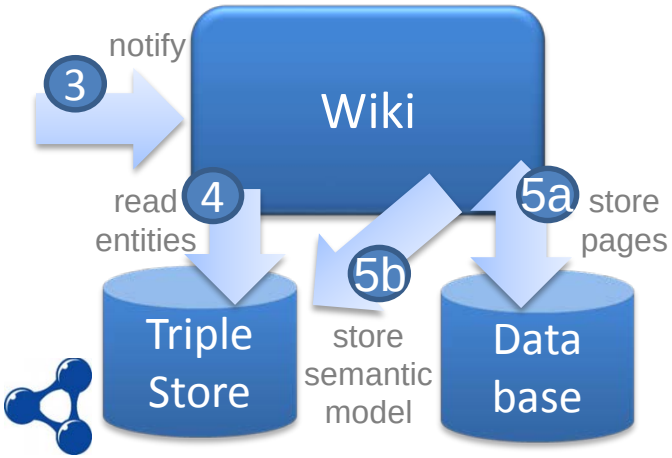
#SemTechBiz

SEMANTIC TECHNOLOGY & BUSINESS CONFERENCE

Re-architect Semantic Wiki

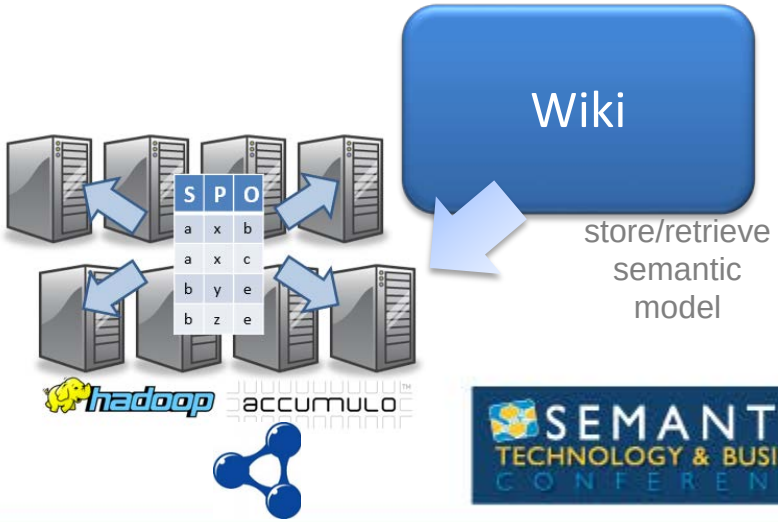
Before

- Read entities from TS and create pages for each
- Write wiki pages to DB
- Write semantic model to TS



After

- No reading of entities to create pages; use entities in place
- No wiki page DB
- Semantic model is only model
  - Wiki 'pages' are just a rendering of the semantic model



#SemTechBiz

SEMANTIC TECHNOLOGY & BUSINESS CONFERENCE

Performance Results

Ingest & Query Performance

Ingest Performance

Date	# docs	# nodes	# map-pers	Initial store size (triples)	time in secs	time in h:m:s	docs/ sec	docs/day (projected)	Triples	Triples / doc
10/24/2013	1,745	1	1	0	1621	27mins, 01sec	1.08	93,312	633,879	363
11/03/2013	25,877	3	?	0	6752	1hrs, 52mins, 32sec	3.83	330,912	8,172,200	316
12/18/2013	18,502	8	8	0	3950	1hr, 5mins, 50 sec	4.68	404,352	5,677,034	306
12/18/2013	18,502	8	18	2.2 billion	2840	47min, 20sec	6.51	562,464	5,677,034	306
12/19/2013	9,209	5	30	0	851	14min, 11sec	10.82	934,848	2,817,377	306
12/19/2013	9,209	5	30	2.2 billion	742	12min, 22sec	12.41	1,072,224	2,817,377	306

#SemTechBiz

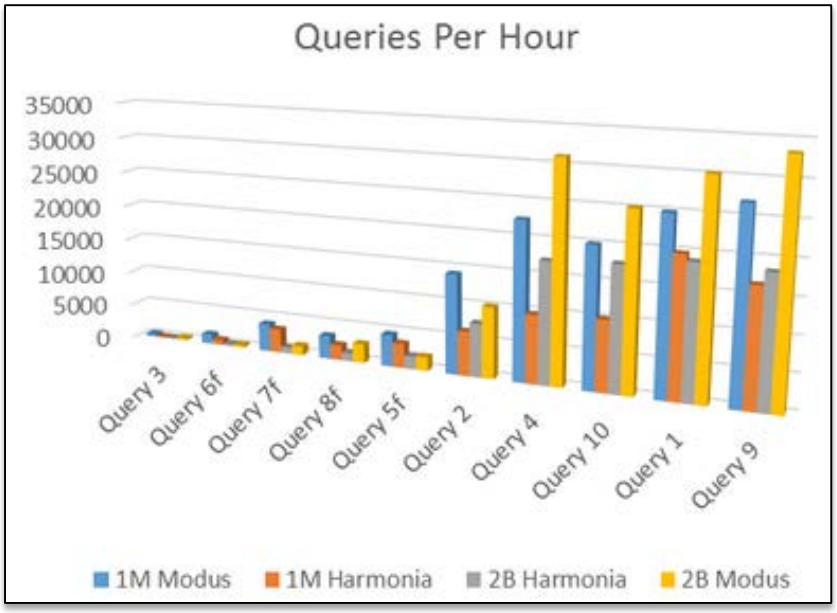
SEMANTIC TECHNOLOGY & BUSINESS CONFERENCE

Query Performance: Benchmark

- Developed our own benchmark
  - A Query Set
    - 14 queries
    - Mimics expected wiki queries
    - Includes text search
  - A Data Generator
    - Generates random data of requested size
    - Generates same data given same seed
    - E.g., Generate:
      - 1M triple dataset using seed 0
      - 2B triple dataset using seed 0

Query Performance: Results

- Over 2 clusters; 2 data sizes



Queries per Hour

	1M Modus	1M Harmonia	2B Harmonia	2B Modus
Query 3	377	116	119	296
Query 6f	1,367	713	286	332
Query 7f	4,003	3,412	852	1,317
Query 8f	3,319	2,155	1,185	2,713
Query 5f	4,590	3,488	1,837	1,967
Query 2	14,255	6,328	7,653	10,185
Query 4	22,577	9,770	17,320	31,213
Query 10	20,054	10,148	17,774	25,250
Query 1	25,010	19,876	18,976	30,276
Query 9	27,011	16,725	18,534	33,410

#SemTechBiz

SEMANTIC TECHNOLOGY & BUSINESS CONFERENCE

Summary: Architecture Changes for Scaling

Before

- Ingest Service:
  - Entity extraction from text by single JVM Web Service
- XWiki with Extensions
  - Wiki pages info stored in PostgreSQL Database
  - Semantic info related to each page stored in triple store
- Triple Store
  - Conventional single node

After

- Ingest Service:
  - Entity extraction from text by Hadoop® MapReduce
- Custom Wiki (from ground up)
  - No PostgreSQL Database
  - Wiki pages rendered directly from triples
- Triple Store:
  - Accumulo™ Cluster based (Rya)

References

1. Apache™ Hadoop®. http://hadoop.apache.org/
2. SPARQL in the Cloud using Rya. Roshan Punnoose, Adina Crainiceanu, David Rapp. http://www.usna.edu/Users/cs/adina/research/Rya_ISjournal2013.pdf
3. Apache™ Accumulo™. https://accumulo.apache.org/

Questions?