

# MarkLogic

Combining Semantic, Relational and Document NoSQL worlds, and doing something useful with it too

Emil Zegers, Systems Engineer, MarkLogic Corporation

© COPYRIGHT 2013 MARKLOGIC CORPORATION. ALL RIGHTS RESERVED.



#### We Are The New Generation Database

#### Hierarchical Era

For your application data!

 Application- and hardware-specific



#### **Relational Era**

"For all your structured data!"

- Normalized, tabular model
- Applicationindependent query
- User control

#### ORACLE"

#### **Any Structure Era**

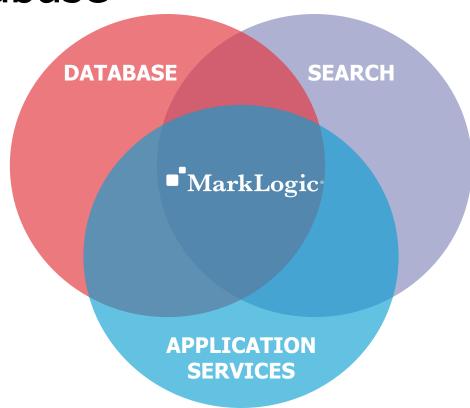
"For all your data!"

- Schema-agnostic
- Massive scale
- Query and search
- Analytics
- Application services
- Faster time-to-results



The Only Enterprise NoSQL Database

- Search & Query
- ACID Transactions
- High Availability / Disaster Recovery
- Replication
- Government-grade Security
- Elasticity
- Cloud Deployment
- Hadoop for Storage & Compute
- Semantics



#### How Does it Work?



Schema-agnostic design



Real-time indexing for search and query



Reverse queries for event processing and alerting



Scale-out shared-nothing cluster topology



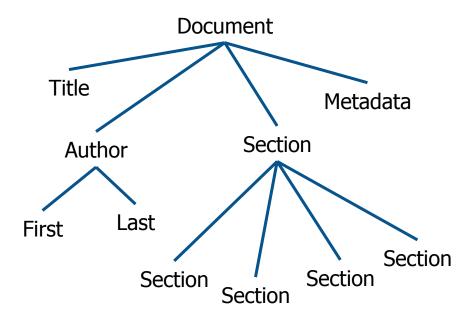
Analytics and Visualization

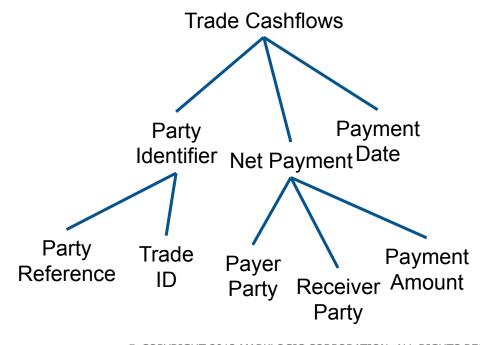


**Transactions** 

#### Hierarchical Data Model

- MarkLogic Server is a document-centric database
  - Supports any-structured data via hierarchical data model
  - Stores compressed binary trees





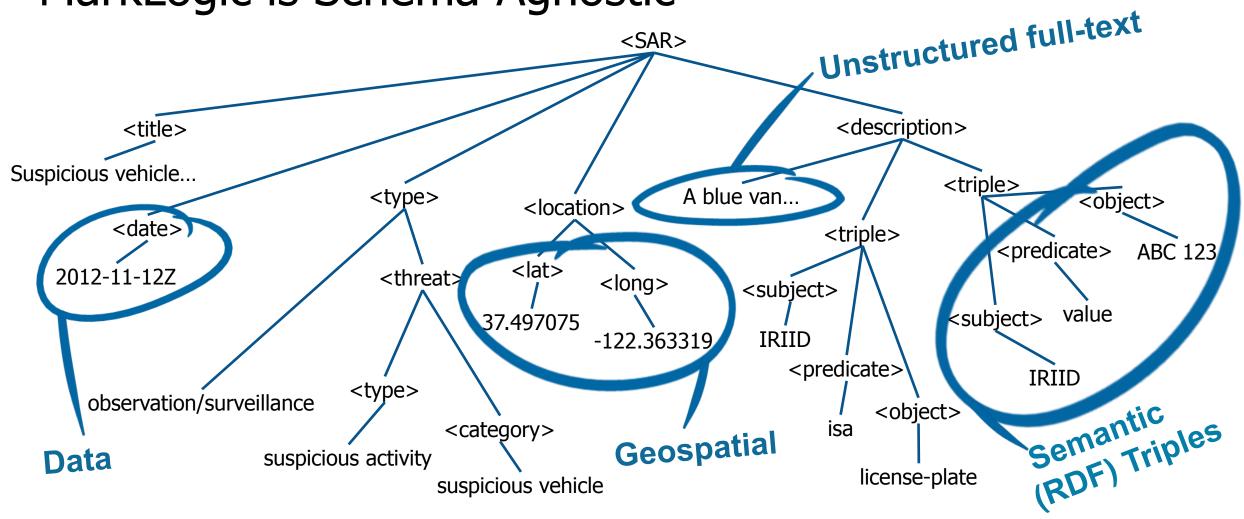


# MarkLogic is Schema-Agnostic

JSON and XML are self-describing:

```
<SAR>
 <titleSuspicious vehicle near airpox/ttitle>
  <date> 2012-11-12Z</date>
  <type> observation/surveillan/de/pe>
 <threat>
   <type> suspicious activity </type>
   <category>suspicious vehi&leategory>
  </threat>
  <location>
   <lat>37.497075
   <long> -122.363319</long>
 </location>
  <descriptionA blue van with license plate ABC 123 was observed parked behind the airport sign...</pre>
   <triplexsubject>IRIIDX/subjectxpredicatexisa/predicatexobject>license-plat/eobjectx/triple>
   <triplexsubject>IRIIX/subjectxpredicatexaluex/predicatexobject>ABC 123</objectx/triple>
  </description>
</SAR>
```

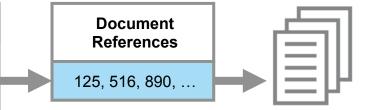
MarkLogic is Schema-Agnostic



#### Universal Index

Term	Term List
"blue"	123, 125, 129, 152, 344, 491,
"van"	123, 125, 126, 129, 130, 152,
"observed"	125, 152, 516, 522, 765, 890,
"blue van"	123, 125, 126, 129, 130, 152,
STEM "observe"	125, 152, 516, 522, 765, 890,
<sar></sar>	
<sar>/<location></location></sar>	
<threat>/<category></category></threat>	
<type>suspicious activity</type>	<b></b>
<date>2012-11-12Z</date>	
Collection:Vetted	<b></b>
Role:Analyst + Action:Read	<b></b>

# Which vetted SAR reports contain the phrase **blue van**?



#### MarkLogic indexes...

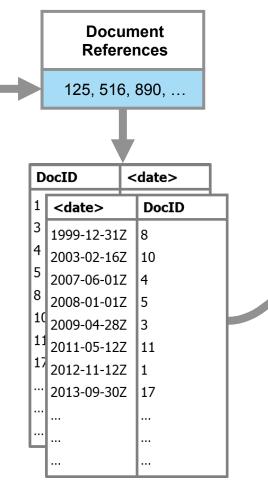
- Words
- Phrases
- Stemmed words and phrases
- Structure
- Data
- Collections
- Security Permissions



# Range Index

Term	Term List
"blue"	123, 125, 129, 152, 344, 491,
"van"	123, 125, 126, 129, 130, 152,
"observed"	125, 152, 516, 522, 765, 890,
"blue van"	123, 125, 126, 129, 130, 152,
STEM "observe"	125, 152, 516, 522, 765, 890,
<sar></sar>	
<sar>/<location></location></sar>	
<threat>/<category></category></threat>	
<type>suspicious activity</type>	
<date>2012-11-12Z</date>	
Collection: veited	
Role:Analyst + Action:Read	

Which vetted SAR reports containing the phrase **blue** van were submitted before 2013?



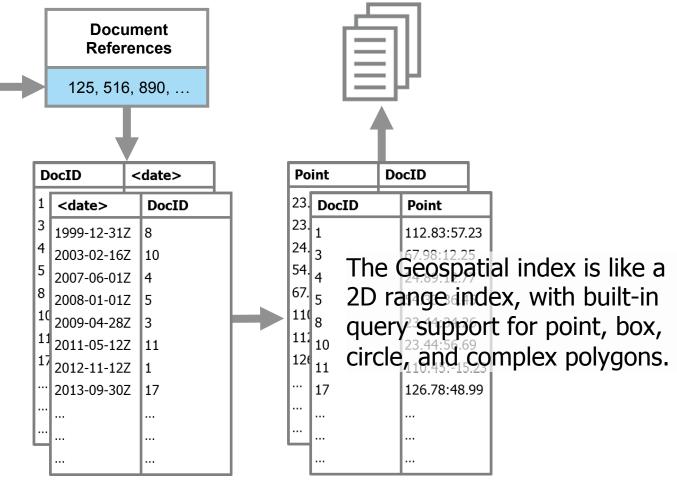
Range indexes map document IDs to values, and vice-versa in a compact in-memory representation.



# Geospatial Index

**Term List Term** "blue" 123, 125, 129, 152, 344, 491, ... "van" 123, 125, 126, 129, 130, 152, ... "observed" 125, 152, 516, 522, 765, 890, ... "blue van" 123, 125, 126, 129, 130, 152, ... STEM "observe" 125, 152, 516, 522, 765, 890, ... <SAR> <SAR>/<location> <threat>/<category> <type>suspicious activity</type> <date>2012-11-12Z</date> Collection: Vetted Role:Analyst + Action:Read ...

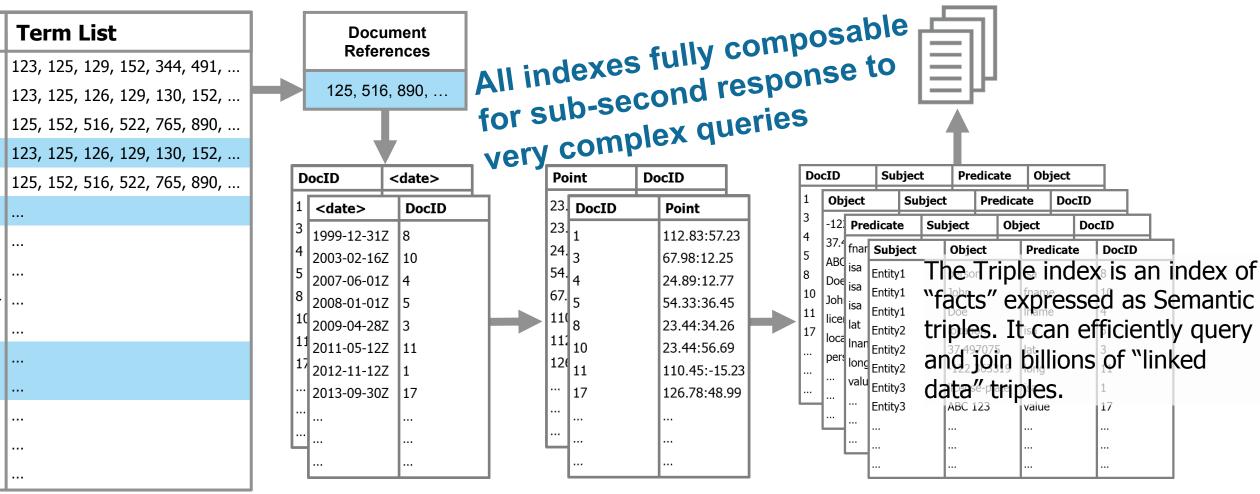
Which vetted SAR reports about a blue van from before 2013 refer to a location near the airport?



#### "MarkLogic

# Triple Index

Which vetted SARs about a blue van from before 2013 with this location refer to partial plate **ABC**?

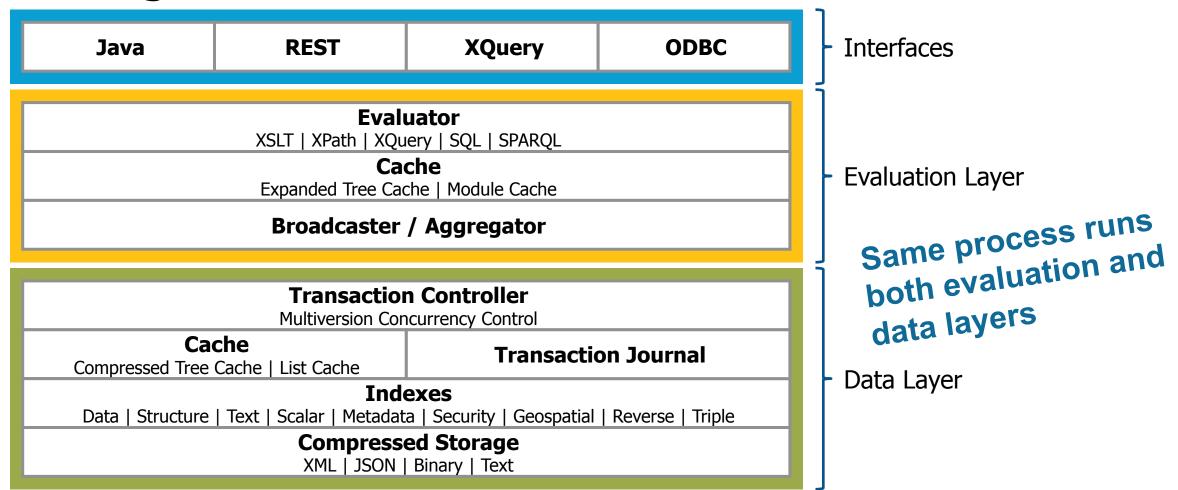


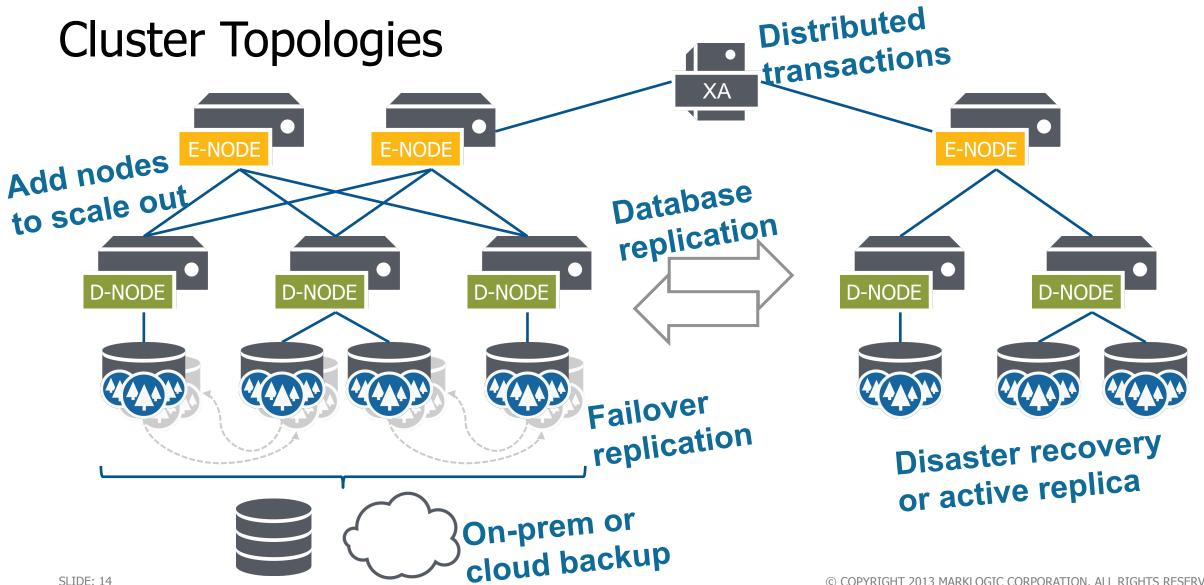
#### Reverse Indexes

- Queries can be serialized as XML and indexed, just like documents
  - word-query("blue van")
    - → <word-query><text>blue van</text></word-query>
- MarkLogic's Alerting feature can
  - Index these queries
  - For a given data document, quickly find all possible matching queries
  - Narrow down these candidates to exact matches using a unified expression tree

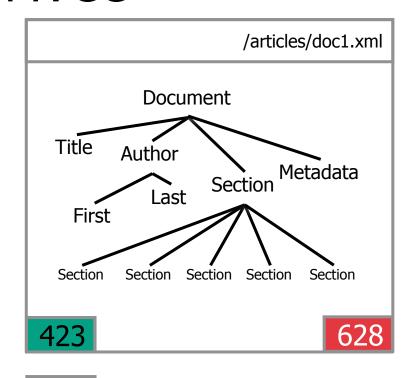
#### "MarkLogic

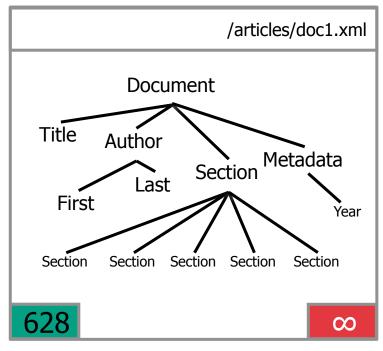
# MarkLogic Architecture





#### MVCC





#### **MVCC Benefits:**

- ACID transactions
- Zero-latency search indexing
- High throughput
- Lock-free reads
- Point-in-time query
- Fast database rollback
- Enables serial writes

Deletion Timestamp

## Enterprise NoSQL

Same trusted, enterprise-class features you find in a RDBMS

#### **Trusted**

Enterprise-ready and secure for mission-critical apps



ACID Transactions



XA Distributed Transactions



Common Criteria Security Certification





Configuration Management







Backup/ Restore



Point-intime Recovery





Journal Archiving



Database Rollback

# DATA



# DATAF

# INFORMATION



# DATAH CONTEXT INFORMATION

# CONTEXT

**DOCUMENTS** 

**DOMAIN** 

WORLD AT LARGE

# Context from the World at Large

#### Linked Open Data

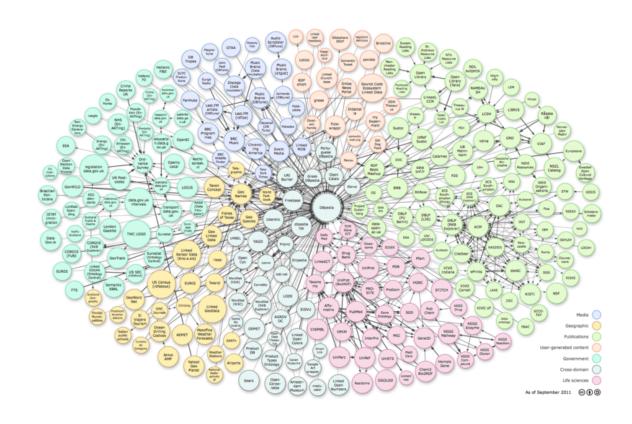
- Facts that are freely available
- In a form that's easily consumed

DBpedia (wikipedia as structured information)

- Einstein was born in Germany
- Ireland's currency is the Euro

#### GeoNames

- Doha is the capital of Qatar
- Doha has these lat/long coordinates



"Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

#### Context from Domain

Like Open Data, but domain specific

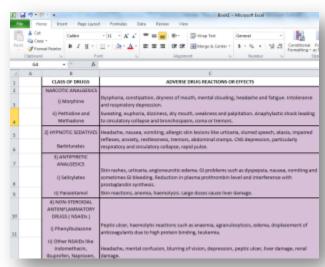
- Might be proprietary within a company
- Or shared across an industry
- Includes data and ontologies

#### Some Examples

- A bank's proprietary reference data
- A pharmaceutical company's drug ontology
- An industry-wide ontology such as FIBO

#### **Proprietary Semantic Facts**

(Facts and Taxonomies in your organization or industry)



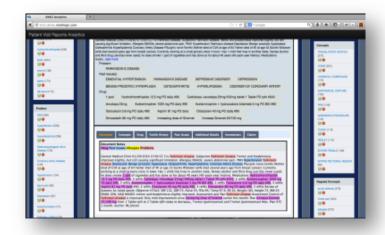
#### Context from Documents

#### Document metadata

Ex: Categories, author, publish date, source

#### Facts in free-flowing text

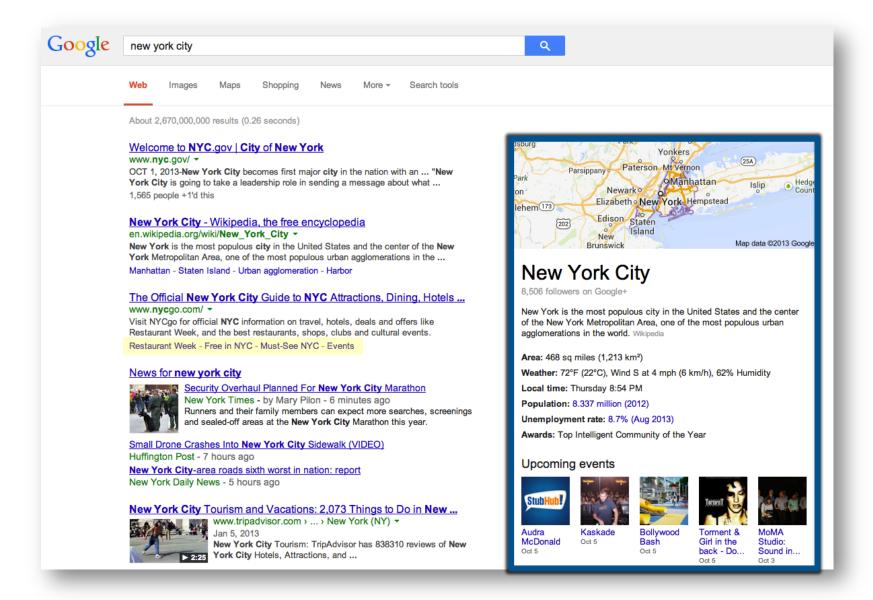
- Entities: this document mentions the person Richard Nixon, the product Advil, the company IBM
- Events: this document says that Nixon went to China, John Smith met Jane Doe, Barclays acquired Lehman Brothers
- Found automatically or provided at authoring time



# Documents can contain triples <article> <meta> <title>Man bites dog</title> <sem:triple> <sem:subject>http://example.org/news/42</sem:subject> <sem:predicate>http://example.org/published</sem:predicate> <sem:object>2013-09-10</sem:object> </sem:triple> ...



# What's behind this?





### Semantics: A New Way to Organize Data

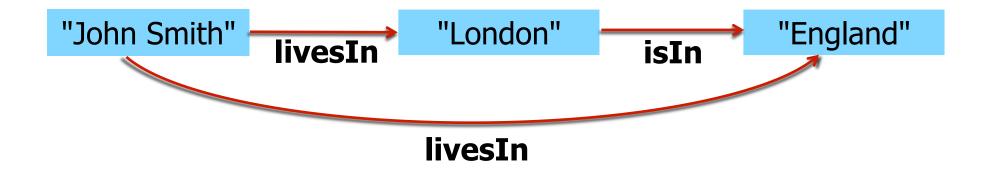
**Data** is stored in Triples, expressed as: **Subject** : **Predicate** : **Object** 

John Smith: livesIn: London

London : isIn : England

Rules tell us something about the triples:

If (A livesIn X) AND (X isIn Y) then (A livesIn Y)

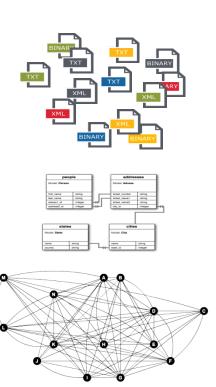


#### "MarkLogic

### Semantics: A New Way to Search Data

- Document search:
   Search full-text and structure
- Database query: Query data (dates, numbers, strings) and ranges
- Semantics:
  Query facts and relationships from many sources

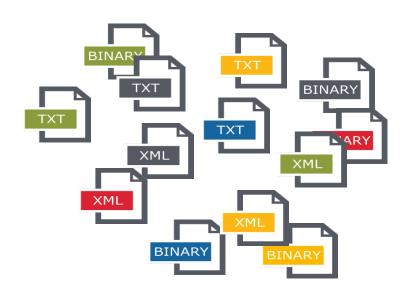
**SPARQL** is a language designed to query triples. It looks a bit like SQL.

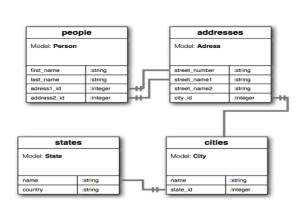


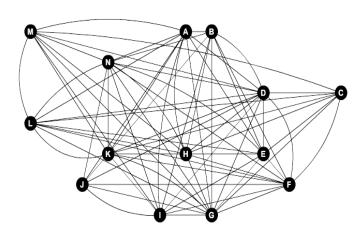
### **Executing SPARQL**

```
sem:sparql("
  prefix : <http://example.org/kennedy/>
  select * {
    ?person :first-name ?first;
      :last-name ?last;
      :alma-mater [:ivy-league :true]
 map:entry("first", "John"),
  cts:collection-query("mycollection")
```

# MarkLogic + MarkLogic Semantics: A New Way to Understand and Discover Information







Document Store + Data Store + Triple Store



### **Dynamic Semantic Publishing**

# BBC Sports The Challenge

- Size and Complexity:
- # of athletes
- # of teams
- # of assets (match reports, statistics, etc.)
- # of relations (facts)

#### **Goals**

- Rich user experience
  - See information in context
  - Personalize content
  - Easy navigation
  - Intelligently serve ads (outside of UK)
- Manageable
  - Static pages? Too many, changing too fast
- Limited number of journalists
  - Automate as much as possible

### **Dynamic Semantic Publishing**

#### A Solution

#### **XML Database**

- Store, manage documents
  - Stories
  - Blogs
  - Feeds
  - Profiles
- Store, manage values
  - Statistics
- Full-Text search
- Performance, scalability
- Robustness

#### **Triple Store**

- Metadata about documents
  - Tagged by journalists
  - Added (semi-)automatically
  - Inferred
- Facts reported by journalists
- Linked Open Data for real-world facts

#### "MarkLogic



#### **Facts into actions:**

Include a story about Lampard

Include in the list of recent matches

Include in the League Table

#### "MarkLogic

# MARKLOGIC SEMANTICS



#### **Triple Store**

Enterprise ready & horizontally-scalable. Store RDF triples with documents and data



#### Triple Index

In addition to data, structure, text, scalar, metadata, security, and geospatial



#### **SPARQL**

Industry-standard query language: compose with data, text, geo & other indexes

### Triple Index

- 3 triple orders
- Cached for performance
- Works seamlessly with other indexes
- Security
- 150 bytes per triple on disk
- Billions of triples per host
  - Scaling out horizontally

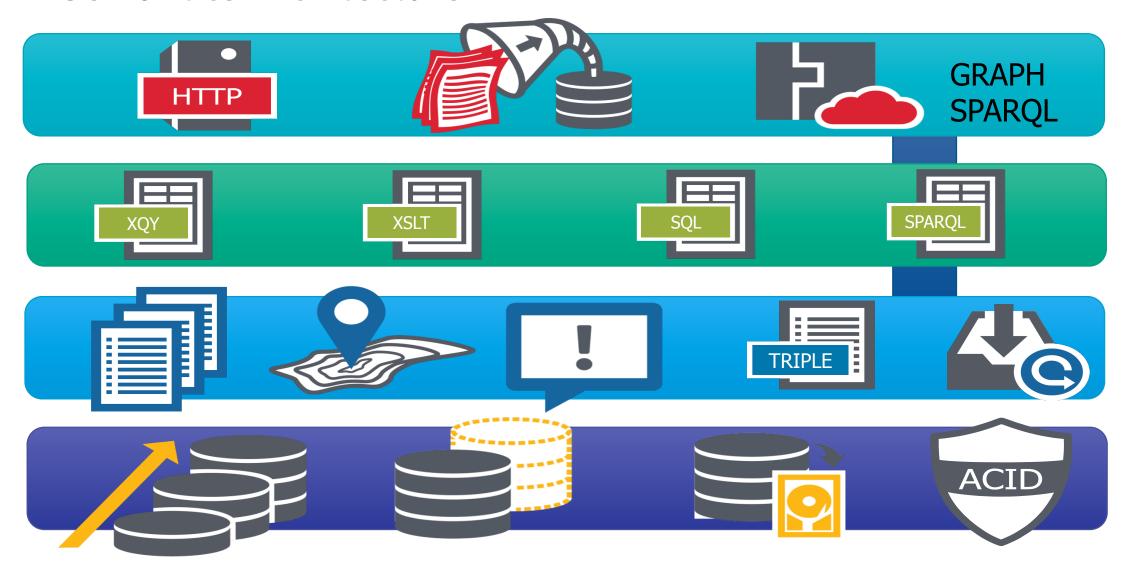








#### **Semantics Architecture**



# SEMANTICS IN ACTION

### Demo time

- MarkMail
   http://markmail.org
- Newz
   <a href="http://newz.nl/?lang=en">http://newz.nl/?lang=en</a>
   <a href="http://www.youtube.com/watch?v=Mf8Lb1Cmhus">http://www.youtube.com/watch?v=Mf8Lb1Cmhus</a>
- Gospel Topic Explore <u>http://cdev.lds.org/semantic</u>
- Combining Semantic with Document search and migrating RDBMS data to the MarkLogic NoSQL DB

https://github.com/mustard57/SYTYCD

http://www.youtube.com/watch?v=4cUqMzsu0F4

()

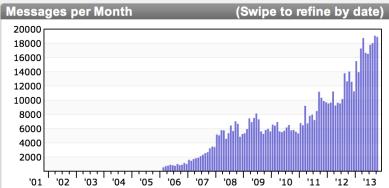
(2)



■ MarkLogic







What List?	View more
org.apache.hadoop.core-dev	83,180
org.apache.hadoop.hive-dev	67,340
org.apache.hadoop.core-user	46,813
Who Sent It? View more Any	Attachments? View more

Who Sent It?	View more	Any Attachments?	View more
Apache Jenkins Server	20,431	png	851
Hadoop QA (JIRA)	11,574	jpg	352
Hudson (JIRA)	10,345	txt	259

1 to 10 of about 661576 👄 🖙 🕥 Sort by Relevance

#### Re: [ANN] HBase 0.20.0-alpha available for download

Is there a requirement for **hadoop** 0.20 for HBase 0.20? Jun 17, 2009 - jason hadoop - org.apache.hadoop.hbase-user

#### re about translating **Hadoop**: The Definitive Guide into chines...

So am I, although i'm just a novice to hadoop. I worked on a project based on the Zookeeper the... ... when I go deep into hadoop. Maybe I can help with chapter 11 Pig. I am now pig's committer. Jan 9, 2010 - hadoop - org.apache.hadoop.general (2)

#### Re: Running hadoop in standalone mode using Eclipse

...JAVA System Property hadoop.log.dir to something Jun 17, 2009 - jason hadoop - org.apache.hadoop.core-dev

#### Re: Hadoop 20.2 release compability

required to execute the same on any current hadoop 0.20.20x releases. Both the Subject: **Hadoop** 20.2 release compability Nov 3, 2011 - bejo...@gmail.com - org.apache.hadoop.core-user

#### Re: Re: Re: Regarding "Hadoop multi cluster" set-up

...0, depending on your hadoop version working in this case. I am getting the following error in the hadoop-root-datanode-slave log file while running the co...

Feb 7, 2009 - jason hadoop - org.apache.hadoop.core-user

#### Re: Hadoop Installation Mappers setting

Other services include non hadoop services as well as hadoop daemons. Subject: Hadoop Installation Mappers setting May 23, 2013 - bejo...@gmail.com - org.apache.hadoop.mapreduceuser

#### Re: Is there any performance issue with Jrockit JVM for Hado...

... lock with profiling **Hadoop** with SUN JVM? Is there any tricky or special setting for Jrockit vm on **Hadoop**? May 9, 2009 - JQ Hadoop - org.apache.hadoop.core-user

#### Re: incremental loads into hadoop

Sam Your understanding is right, hadoop definitely works great with largeGiga, Tera or Peta bytes. Mostly when said hadoop process tera bytes of data. It Subject: Re: incremental loads into hadoon

#### **Apache Hadoop**



Apache Hadoop is a software framework that supports data-intensive distributed applications under a free license. It enables applications to work with thousands of computational independent computers and petabytes of data. Hadoop was derived from Google's MapReduce and Google File System (GFS) papers. Hadoop is a toplevel Apache project being built and used by a global community of contributors, written in the Java programming language.

#### Releases

• Stable: 1.0.4 download

Unstable: 2.0.2-alpha download

#### **Key Personnel**

- Doug Cutting
- Raghu Angadi
- Owen O'Malley

#### **Related Projects**

- Apache Zookeeper
- Apache Pig

ෂෙ 🔝

Search

Sign In or Sign Up (Why?)



Is there a requirement for **hadoop** 0.20 for HBase 0.20?
Jun 17, 2009 - jason hadoop - org.apache.hadoop.hbase-user

#### re about translating Hadoop: The Definitive Guide into chines...

So am I, although i'm just a novice to **hadoop**. I worked on a project based on the Zookeeper the... ... when I go deep into **hadoop**. Maybe I can help with chapter 11 Pig. I am now pig's committer.

Jan 9, 2010 - hadoop - org.apache.hadoop.general

#### Re: Running hadoop in standalone mode using Eclipse

...JAVA System Property **hadoop**.log.dir to something
Jun 17, 2009 - jason hadoop - org.apache.hadoop.core-dev

#### Re: Hadoop 20.2 release compability

required to execute the same on any current **hadoop** 0.20.20x releases. Both the Subject: **Hadoop** 20.2 release compability

Nov 3, 2011 - bejo...@gmail.com - org.apache.hadoop.core-user

#### Re: Re: Re: Regarding "Hadoop multi cluster" set-up

...0, depending on your **hadoop** version working in this case. I am getting the following error in the **hadoop**-root-datanode-slave log file while running the co...

Feb 7, 2009 - jason hadoop - org.apache.hadoop.core-user

#### Re: Hadoop Installation Mappers setting

Other services include non hadoop services as well as hadoop daemons. Subject: Hadoop Installation Mappers setting May 23, 2013 - bejo...@gmail.com - org.apache.hadoop.mapreduceuser

#### Re: Is there any performance issue with Jrockit JVM for Hado...

... lock with profiling **Hadoop** with SUN JVM? Is there any tricky or special setting for Jrockit vm on **Hadoop**?

May 9, 2009 - JQ Hadoop - org.apache.hadoop.core-user

#### Re: incremental loads into hadoop

Sam Your understanding is right, **hadoop** definitely works great with largeGiga,Tera or Peta bytes. Mostly when said **hadoop** process tera bytes of data. It Subject: Re: incremental loads into **hadoop** 

#### **Jason Venner**





(

()

(2)

()

San Francisco

Known for: Pro Hadoop, Apress



r download

Base 0.20?

On wed, Jun 17, 2009 at 1:44 AM, Andrew Purtell <apur...@apache.org> wrote:

Minor correction/addition: "Stargate" is undergoing shared development in two github trees:

http://github.com/macdiesel/stargate/tree/masterhttp://github.com/apurtell/stargate/tree/master

The github network view can show which branch is leading at any point in time:

http://github.com/apurtell/stargate/network

- Andy

From: stack <sta...@duboce.net>

To: hbas...@hadoop.apache.org; core...@hadoop.apache.org

Sent: Tuesday, June 16, 2009 10:23:20 PM

Subject: [ANN] HBase 0.20.0-alpha available for download

An alpha version of HBase 0.20.0 is available for download at:

#### **Apache Hadoop**



Apache Hadoop is a software framework that supports data-intensive distributed applications under a free license. It enables applications to work with thousands of computational independent computers and petabytes of data. Hadoop was derived from Google's MapReduce and Google File System (GFS) papers. Hadoop is a top-level Apache project being built and used by a global community of contributors, written in the Java programming language.

#### Releases

Stable: 1.0.4 download

• Unstable: 2.0.2-alpha download

#### **Key Personnel**

- Doug Cutting
- Raghu Angadi
- Owen O'Malley

#### **Related Projects**

- Apache Zookeeper
- Apache Pig





obama Soeken

Instellingen... Uitloggen

Gerelateerde concepten Personen **▼**8 Helima Obama Kezia Obama Fernando Obama Galindo Gebeurtenissen Thomas More Law Center v. Obama Organisaties **▼**8 Republicans for Obama Obama campaign Obama Works Locaties Ohama

geen resultaten

Verfijnen

Datumfilter
Start:

Eind:

Zoeken

## **GTE** beta

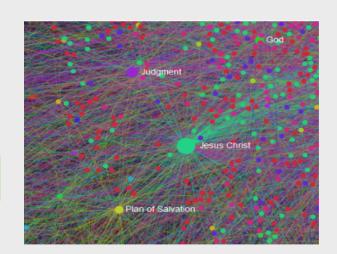
The Gospel Topics Explorer is a concordance of various LDS-related study helps, including the Bible Dictionary, Topical Guide, Guide to the Scriptures, Teaching by Topic and many others.

Relate

Entity enrichment has been added! Now you can see the important entities in your content and link them into the GTE graph.

Learn more »

Get started »



#### **Explore**

Navigate the Gospel Ontology by zooming in anywhere among the 6,600+ connected topics, or click a node for more details. Focus on a given cluster of nodes or scroll back for the bigger picture.

View details »

#### Search

Search more than 11,000 topics, 23,000 relationshps and 67,000 facts by keyword, phrase or topic. Results are ordered by the number of relationships each topic contains and includes direct links to examples of the topic.

View details »

### Find relationships

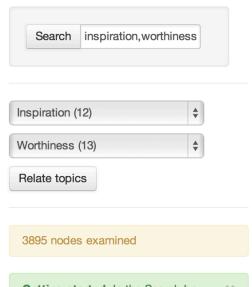
Ever wonder what the relationship between sacrament and revelation is? Or how sacrifice and the atonement are connected?

View details »

#### **Enrich**

Find the topics, people, organizations and roles embedded in content. Just copy and paste some text into the form and the GTE will identify entities of interest.

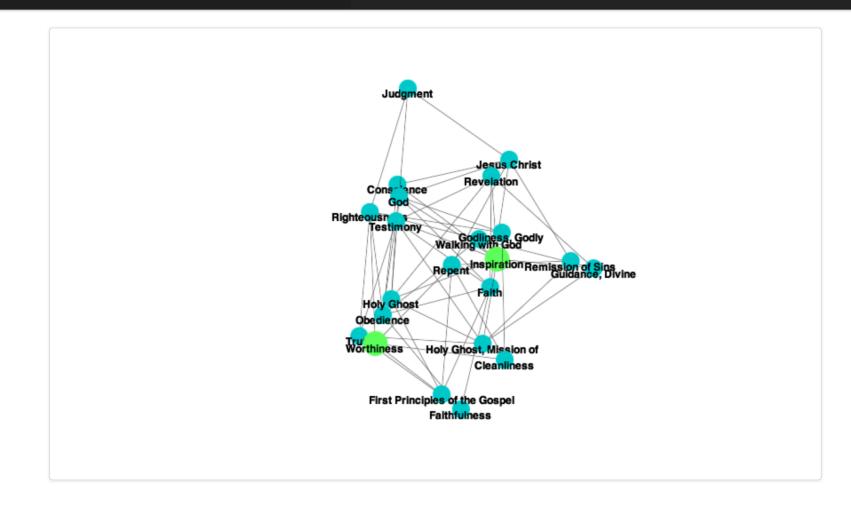
View details »



Getting started: In the Search box above, enter two terms or phrases separated by a comma, then click Search. Select other topics from one or both dropdowns to refine your search.

#### Examples

atonement, joseph smith wealth, misery inspiration, worthiness sacrament, atonement worthiness, mission cleanliness, godliness tea, coffee

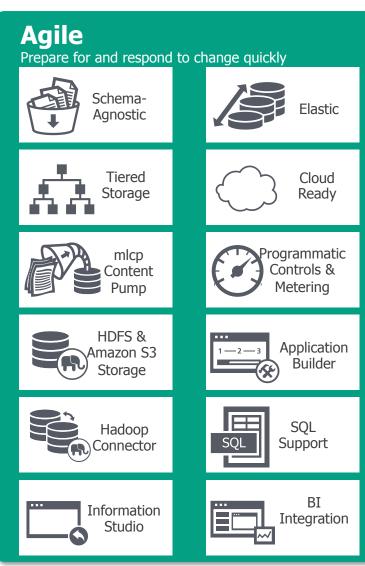


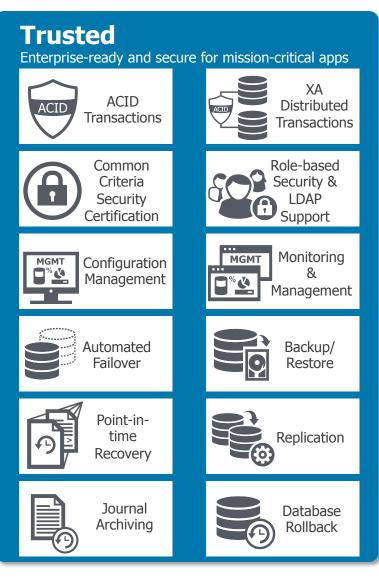


#### Subject Search Results Related Documents **Entity Facts** Load related content NKB Customer: Broussard Search Search: Load related content 1. Ms. Theodora Broussard (Joint Customer) Account: 96302787 (NKB Account) Results 2. Mr. Jeffrey Blow (Joint Customer) Address: /docs/2.html No Results 3. Mr. John Castillo (Joint Customer) Branch: Broussard (NKB Customer) 4. Mr. James Bowers (Joint Customer) Home Phone Number: 01422 518885 5. Mr. Cory Jackson (Joint Customer) Mobile Phone Number: 07838 147787 Address ID: 11 6. Mrs. Yvone Crawford (Joint Customer) Customer ID: 11 7. Mr. Robert Nelson (Joint Customer) Date Of Birth: 1964-12-19 8. Dr. Carole Smoot (Joint Customer) Branch ID: 9 9. Mrs. Joanne Miller (Joint Customer) Initial: A

# Enterprise NoSQL: Powerful, Agile & Trusted









# ANY QUESTIONS?

### **Contact Information**

**Emil Zegers** 

Pre-Sales Consultant Benelux & Nordics

emil.zegers@marklogic.com

+31 619929703

Graadt van Roggenweg 328-334, Blok D 3531AH Utrecht, Netherlands

http://marklogic.com

## Useful links [1]

MarkLogic Developer Community

http://developer.marklogic.com

Inside MarkLogic Server

http://developer.marklogic.com/pubs/architecture/inside-marklogic-server-r7.pdf

MarkLogic Application Development

http://docs.marklogic.com/guide/app-builder/intro

## Useful links [2]

What's New in MarkLogic 7

http://www.marklogic.com/what-is-marklogic/marklogic

MarkLogic MLJS

https://github.com/adamfowleruk/mljs

MarkLogic Blog

http://developer.marklogic.com/blog