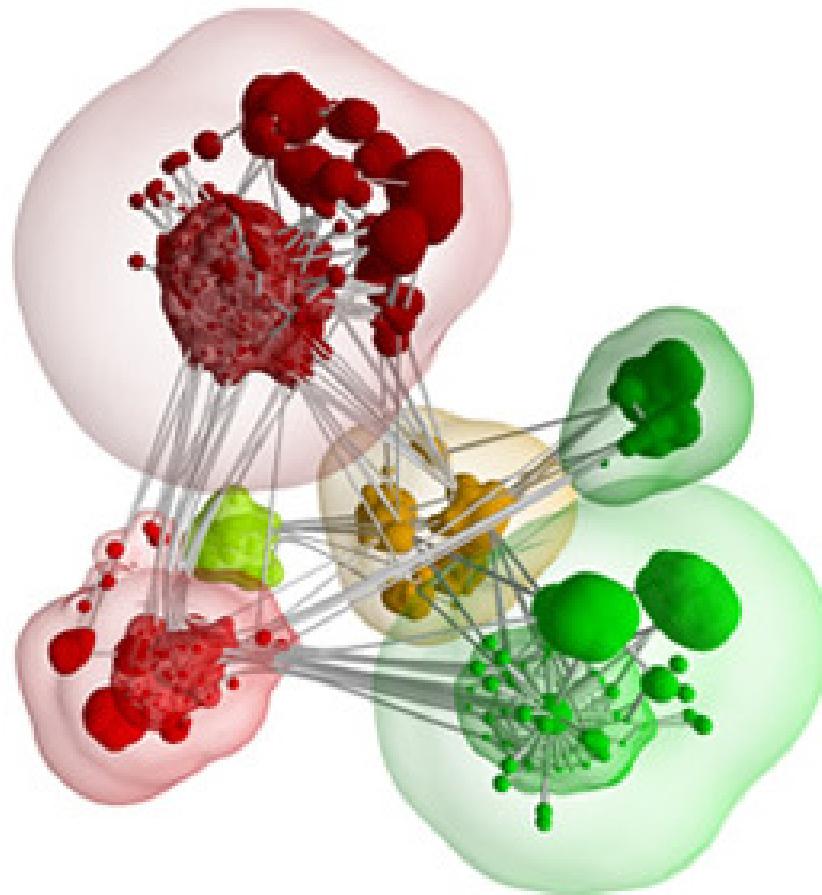




Triple Stores or Nosql in the Enterprise

Jans Aasman, Ph.D.
CEO Franz Inc
Ja@Franz.com



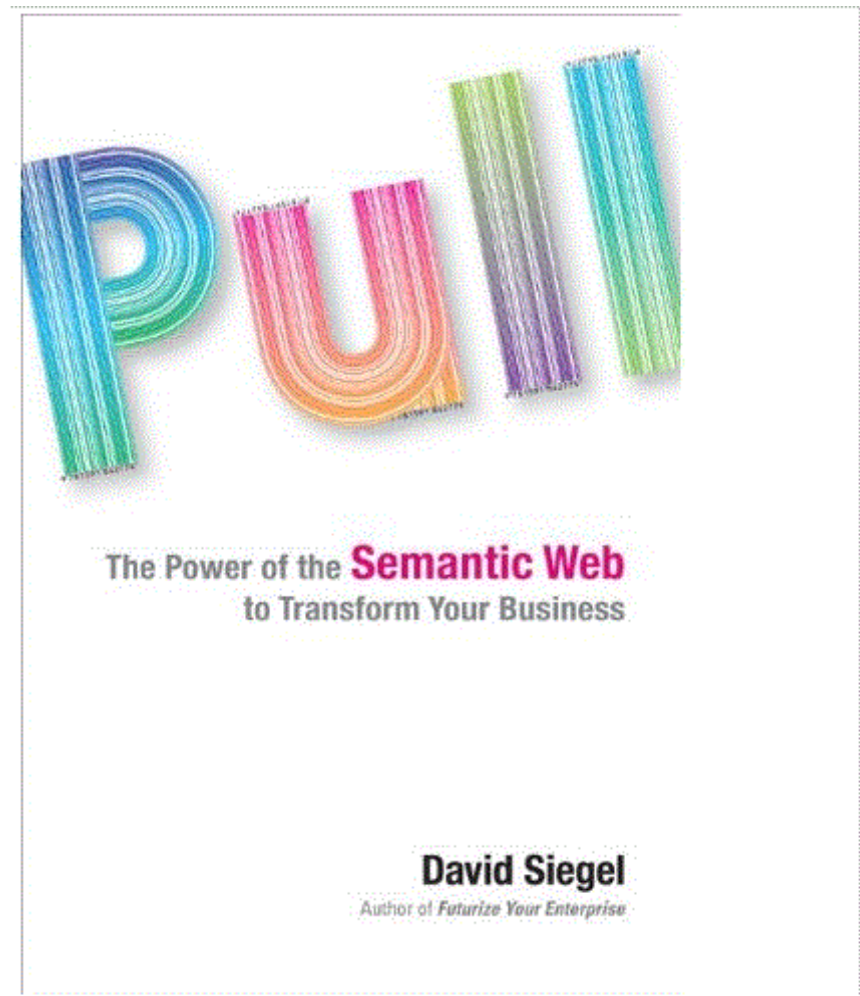
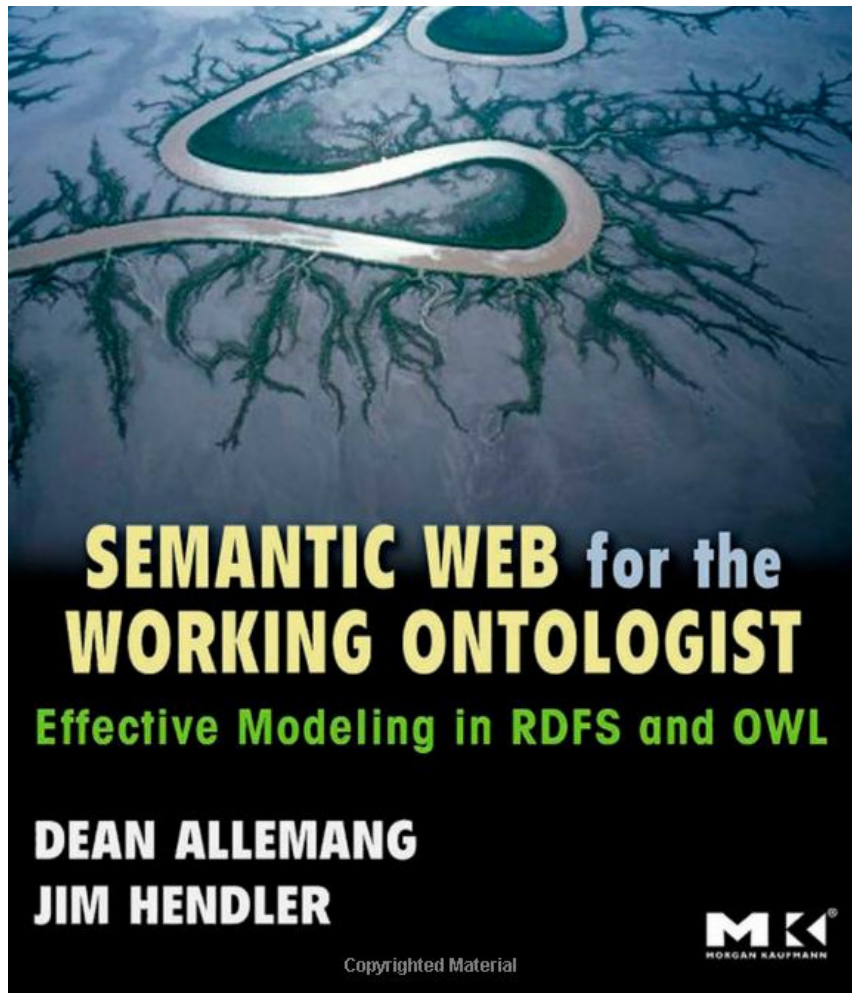


Questions we get every day..

- Can you explain more about triples and meta data?
- What is the difference between a RDB and a triple store?
- What do companies use triple stores for?
- And why do companies work with triple stores?
- Can't I do that with a Nosql database like Hadoop, Bigtable, HBase, Cassandra...
- General requirements for RDB, Triple Stores and Big Data
- What is so special about your triple store...
- Can I scale to a trillion triples and what do I have to do for that?
- And from our lab: our new visual query editor



Can you explain more about triples and meta data?





**What is the difference between
a RDB and a triple store?**



An artist's impression of a db for persons

Table Person

ID	First-Name	Last-Name	Middle-In.	DOB	DOD	PlaceOB	Sex
2	Rose	Fitzgerald	Elizabeth	1890	1995	1	F

Table Spouses

ID1	ID2
2	1

Table to-schools

ID1	SchoolID
2	3

Table Schools

ID	Name
3	Sacred-Heart

Table has-profession

ID1	ProfID
2	3

Table Professions

ID	Name
3	Home-maker

Table Has-Child

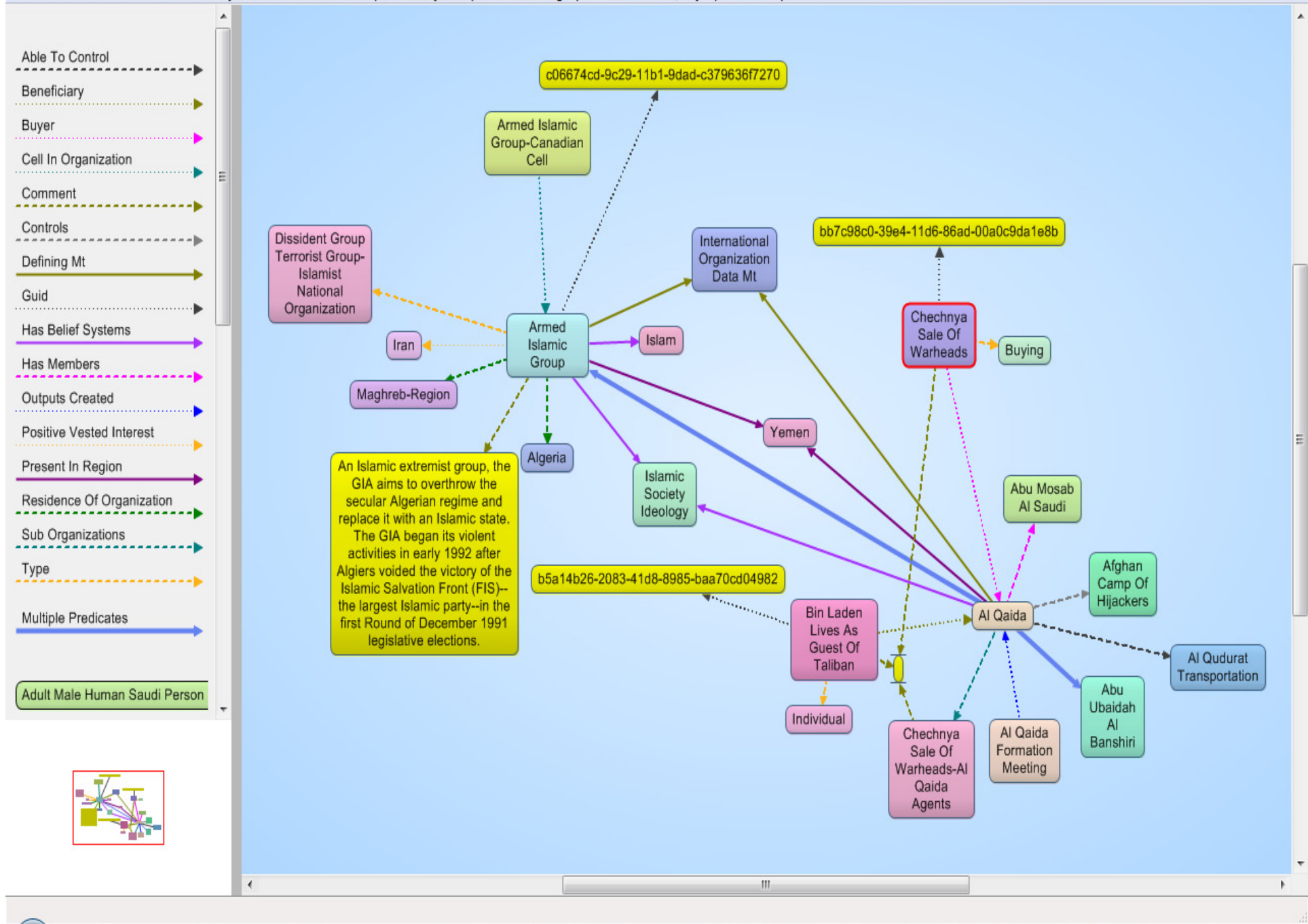
ID1	ID2
2	17
2	15
2	14
2	13

Table Place

ID	Name	State	Longitude	Latitude
1	Boston	MA	42.3	-71.4



person2	type	person
person2	first-name	Rose
person2	middle-initial	Elizabeth
person2	last-name	Fitzgerald
person2	suffix	none
person2	alma-mater	Sacred-Heart-Convent
person2	birth-year	1890
person2	death-year	1995
person2	sex	female
person2	spouse	person1
person2	has-child	person17
person2	has-child	person15
person2	has-child	person13
person2	has-child	person11
person2	has-child	person9
person2	has-child	person7
person2	has-child	person6
person2	has-child	person4
person2	has-child	person3
person2	profession	home-maker



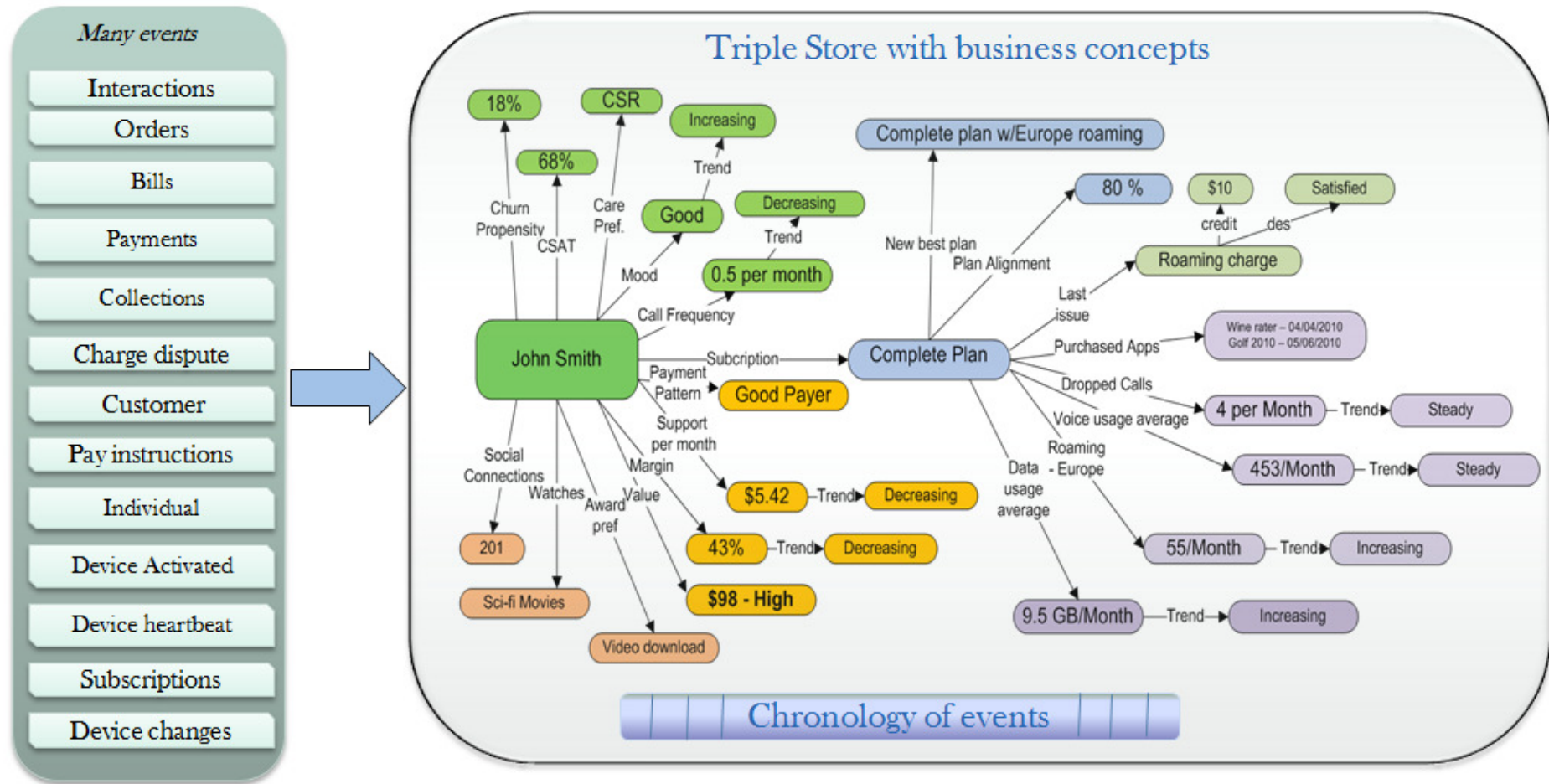


**What do companies use triple
stores for?**

Guided Interaction Advisor Visualization of the solution

See

Events from many source systems are transformed into a set of related business concepts





And why do these companies work with triple stores?

When you need ultimate flexibility

- Modeling knowledge and assets
- Hundreds to thousands of classes with different features
- Everyday new classes and new features
- You work with rules and reasoning

When you need ultimate 'linkability'

- For (ad hoc) integration of databases

When you need pattern recognition and network analysis

- Complex networks of people, companies, products, etc

When you need event processing using geospatial, temporal reasoning and social network analysis combined with flexible metadata



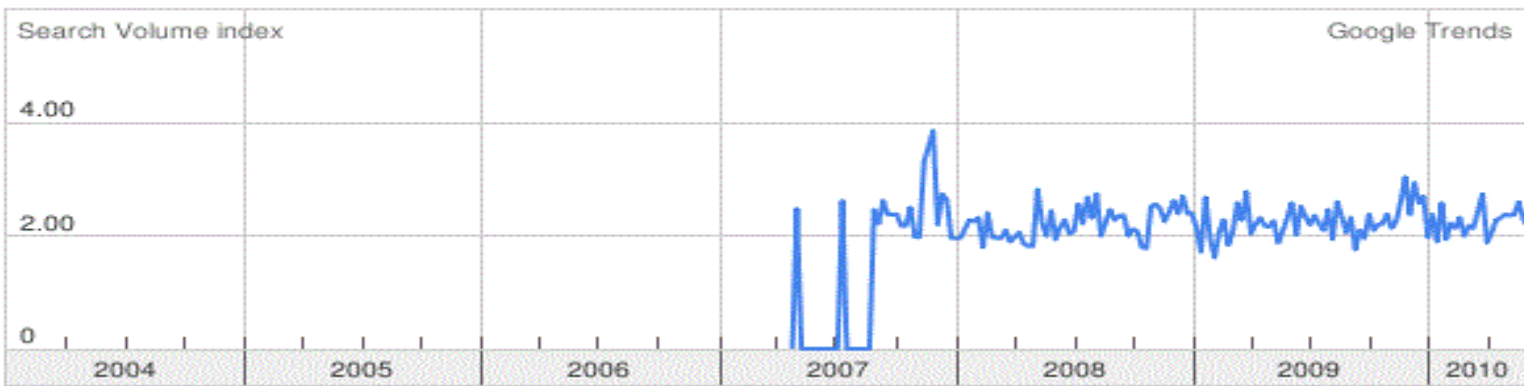
Can't I do that with a Nosql database like Hadoop, HBase, Cassandra...

- HBase is an open-source, distributed database modeled after Google's BigTable and written in Java. It is developed as part of Apache Software Foundation's Hadoop project and runs on top of HDFS (Hadoop Distributed File System), providing BigTable-like capabilities for Hadoop. ...
en.wikipedia.org/wiki/Hbase
- It sacrifices ACID-ness and complex Joins for web scale scalability.

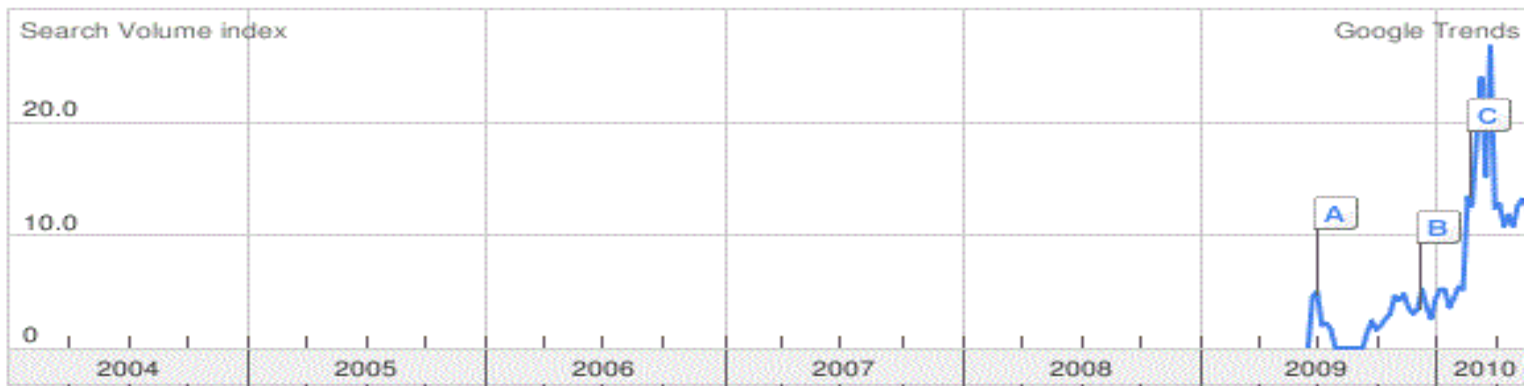
relational database



triple store

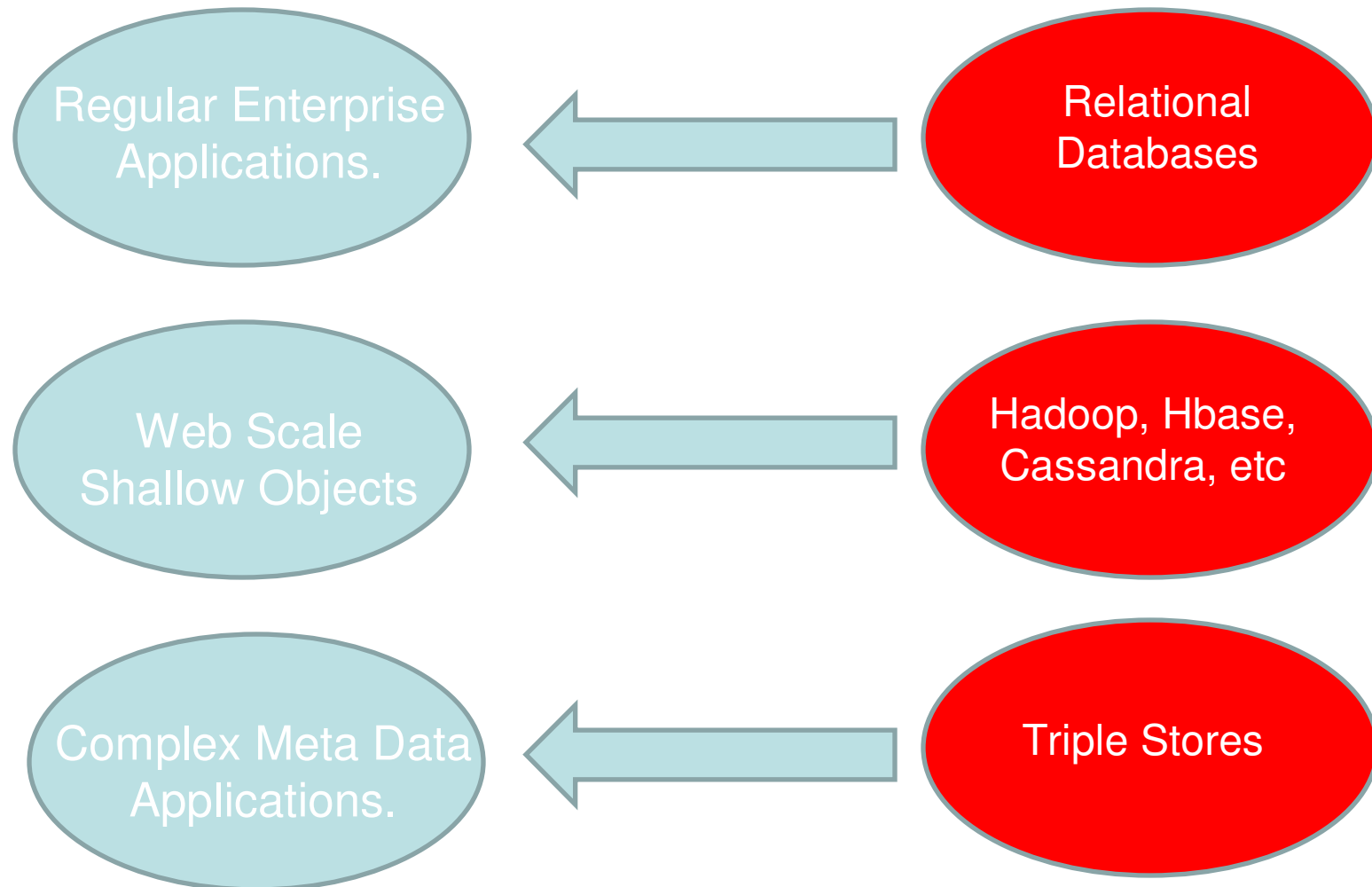


nosql



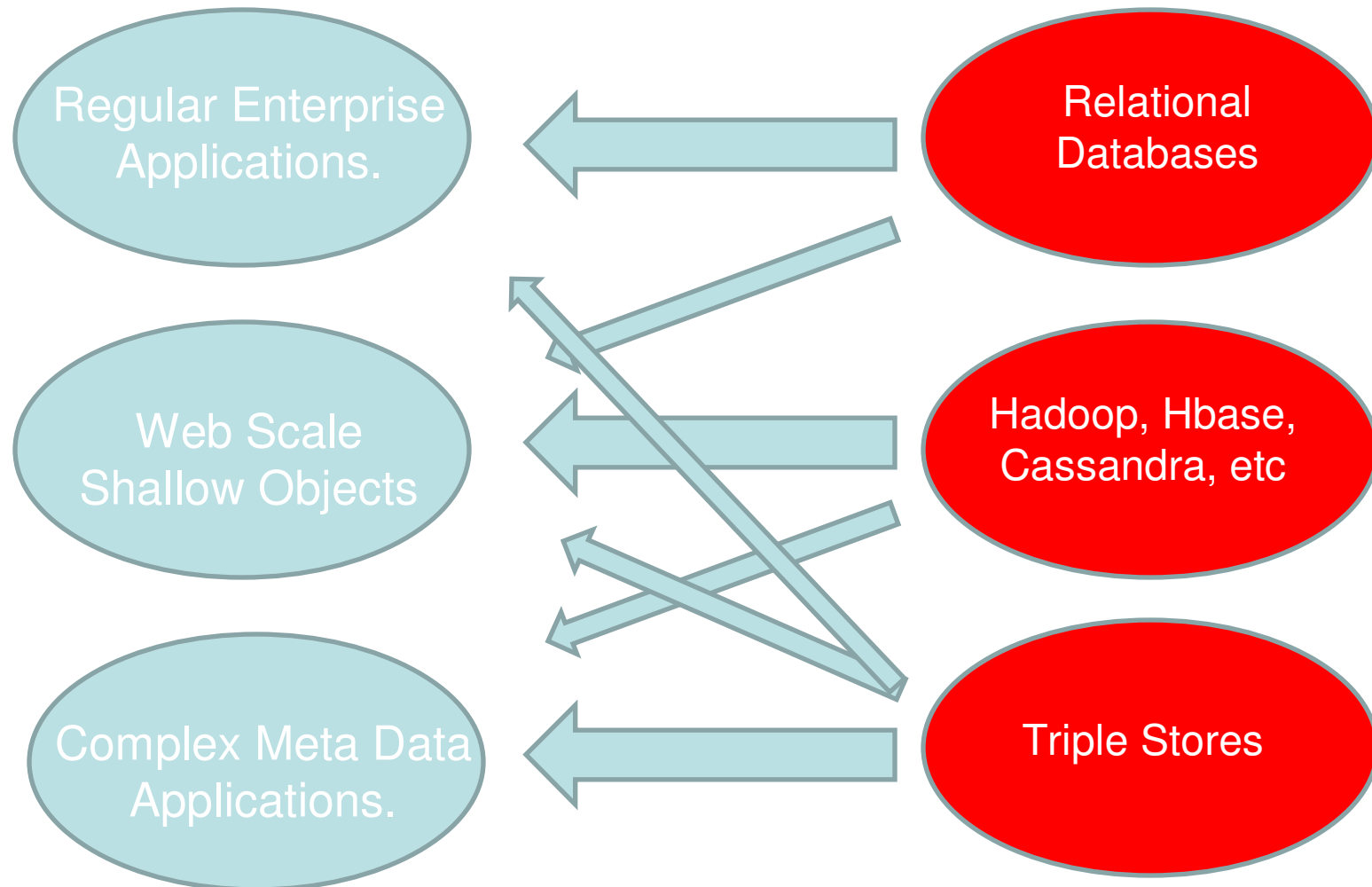


Classes of applications and their natural database.





Classes of applications and their natural database.





Comparing RDB, HBase, Triple Stores

	RDB	HBase	Average Triple Store
ACID	+	-	-
Concurrent and Dynamic	+	+	-
Random Access	+	+	+
Flexibility	-	-	+
High Availability	+	+	-
Very complex graph search	-	-	+
Structured + Unstructured	-	-	+
Scalability	+	+	-



**What is so special about your
triple store...**



AllegroGraph [1]

- Scalable and persistent Triple Store
 - Well, actually: $s \ p \ o \ g + i$
 - That can store billions of triples on a single machine
- Federated
 - Create an abstract store that is a collection of other triple stores. Prolog and SPARQL and Reasoning work transparently against abstract store
- Compliant with standards
 - RDF, RDFS, OWL, SPARQL, Named Graphs, ISO Prolog, OWL-lite reasoning



AllegroGraph [2]

- Relational database efficiency for range queries
 - We support most xml schema types (dates, times, longitudes, latitudes, durations, telephone numbers, etc)
- Spatial database efficiency for geospatial primitives
 - Find elements in bounding boxes as fast as in spatial databases
- Temporal reasoning
 - Reasoning about times and intervals (Allen Logic)
- Social Network Analytics library
 - Find actor degrees and centrality, cliques, group centrality and cohesiveness



A Simple Event Ontology

- A type
 - Meetings, communications event, financial transactions, visit, attack/truce, an insurance claim, a purchase order
 - RDFS++ reasoning
- A list of actors
 - Social Network Analysis
- A place
 - GeoSpatial Reasoning
- A Start-time and possible an end-time
 - Temporal Reasoning
- Anything else that describes the event
 - Goods that changed hands



Activity Recognition

- Our customers use AllegroGraph as an event database with social network analysis and geospatial and temporal reasoning

Find all meetings that happened in November within 5 miles of Berkeley that was attended by the most important person in Jans' friends and friends of friends.

```
(select (?x)
  (ego-group person:jans knows ?group 2)
  (actor-centrality-members ?group knows ?x ?num)
  (q ?event fr:actor ?x)
  (qs ?event rdf:type fr:Meeting)
  (interval-during ?event "2008-11-01" "2008-11-06")
  (geo-box-around geoname:Berkeley ?event 5 miles)
!)
```

SNA
SNA
DB Lookup
RDFS
Temporal
Spatial



AllegroGraph 4.0

- ✓ Transactions
 - ✓ ACID
 - ✓ Commit, Rollback
 - ✓ Check Pointing, Full and fast Recoverability.
- ✓ Transparent/Automatic Indexing of everything
- ✓ Read and Write concurrency
- ✓ Hot Backups
- ✓ Replication and Warm Failover
- ✓ Lucene style freetext indexing
- ✓ Clustering (use all cpus, memory, disks)
- ✓ Duplicate triple removal
- ✓ Garbage collection of deleted triples
- ✓ One click install (well, for the server that works, and edit config file)



Comparing RDB, Hbase, Triple Stores

	RDB	Hbase	Triple Store	AG4.0
ACID	+	-	-	+
Concurrent & Dynamic	+	+	-	+
Random Access	+	+	+	+
Flexibility	-	-	+	+
High Availability	+	+	-	+
Joins / complex graph search	-	-	+	+
Structured + Unstructured	-	-	+	+
Scalability	+	+	-	+



Performance example: LUBM 8000

- Machine: 32 Gig machine, 4 disks, with 4 cores (2000)
- Indices: spogi and posgi
- One step load + all 14 LUBM queries in 2:25 hours,
 - No preprocessing of strings, no indexing phase, no materialization
 - Ready to go query and reason
- LUBM 40,000 now loads and indexes on a 48 Gig machine in 17 hours. (5.5 Billion triples)



However more important

- Customers want fair performance for simultaneous
 - Adding triples
 - Deleting triples
 - Queries



So we encourage adoption of an Events Test Benchmark

- Goal: how good is the database in adding, deleting and queries triples at the same time and after the first X million triples have been added

- A proposed first version of the events test is at

<http://github.com/franzinc/agraph-python/blob/master/stress/events/events>

stress/events/events at master from franzinc's agraph-python - GitHub - Mozilla Firefox

File Edit View History Delicious Bookmarks Tools Help

git http://github.com/franzinc/agraph-python/blob/master/stress/events/events

GCAL Gmail NBA Movies Sudoku Ttext Weather a P WWT nwa IMDB IMDB YouTube BB wheather Franz psf USTA kranten queries h Digg Campo

git stress/events/events at master fr... x git tutorial at master from franzinc's ag... x Python API Tutorial for AllegroGrap... x

github SOCIAL CODING

Home Pricing and Signup Explore GitHub Blog Login

Search GitHub...

franzinc / agraph-python

Watch Fork Download Source 3 2

Source Commits Network (2) Issues (0) Downloads (0) Wiki (1) Graphs Branch: master

Switch Branches (4) Switch Tags (0) Branch List

AllegroGraph Python client
<http://opensource.franz.com>

HTTP Git Read-Only <http://github.com/franzinc/agraph-python.git> This URL has Read-Only access

AG4 Python tutorial update.

BruceDClayton (author)
February 04, 2010

commit 43a3da501611f9a9ecba
tree 94b2b808e6e169fef9ce
parent d76bacd8e649d0ba3e79

agrap-python / stress / events / events

100755 | 864 lines (708 sloc) | 30.155 kb raw blame history

```
1 #!/usr/bin/env python
2 # -*- coding: utf-8 -*-
3
4 ##### BEGIN LICENSE BLOCK #####
```

Done

Presentati... Windows ... stress/eve... Mozilla T... AllegroGr... Microsof... Snipping ... 12:46 PM



Can I do a trillion



Loading



Fast Queries



On the road map

- We do now routinely 20 Billion triples on a big blade machine
- Expect a trillion triples in December
- Problems we are solving
 - Keep it ACID
 - Smart partitioning
 - Smart (Re-)balancing
 - Smart indexing
 - Query pipelines
 - Do (a part of) a query where the data is.
 - Parallel query execution

