

RDF on Cloud Number Nine

An Implementation, Experiment and Thoughts on
Cloud Computing and Semantic Data Management

30.5.2009

Dr. Valentin Zacharias



WIR FORSCHEN FÜR SIE

- Who?
- What?
- How?
- Does it work?
- And now?

Who?



Valentin Zacharias
Deputy Head of WIM @ FZI



Raffael Stein
PhD Student at AIFB @ KIT

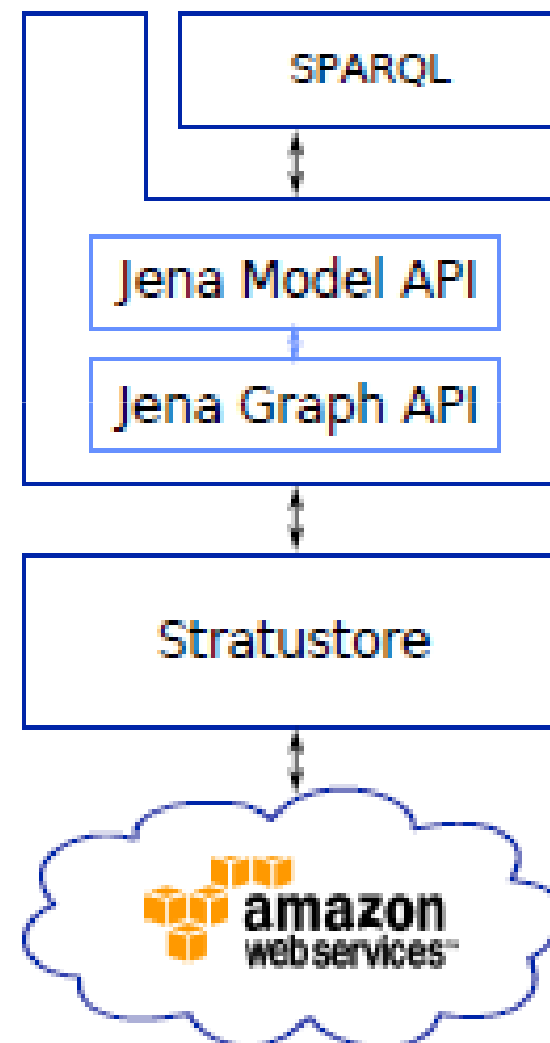
- Build a „worry free triple store service on the cheap“
 - Worry Free:
 - Takes care of backup
 - Scales elastically to large datasets and highly concurrent access
 - Highly available
 - Payment only based on use
 - On the Cheap:
 - Don't start from scratch, build on top an existing ‚DaaS‘



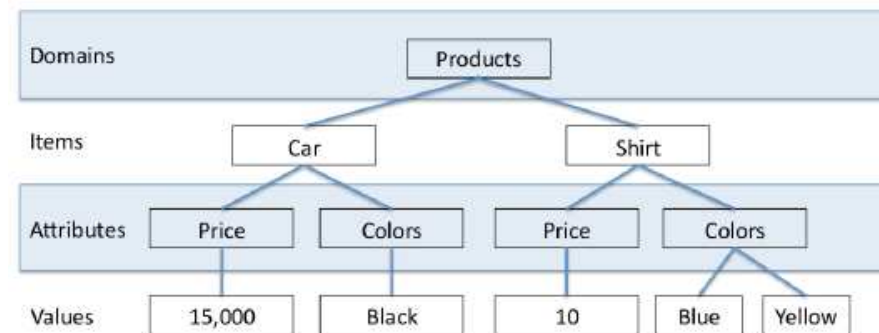
How?

FZI

- Create an implementation of the Jena Graph API that is backed by Amazon's SimpleDB
- SimpleDB
 - Database service by amazon
 - BigTable like, no schema
 - Simple query language without joins



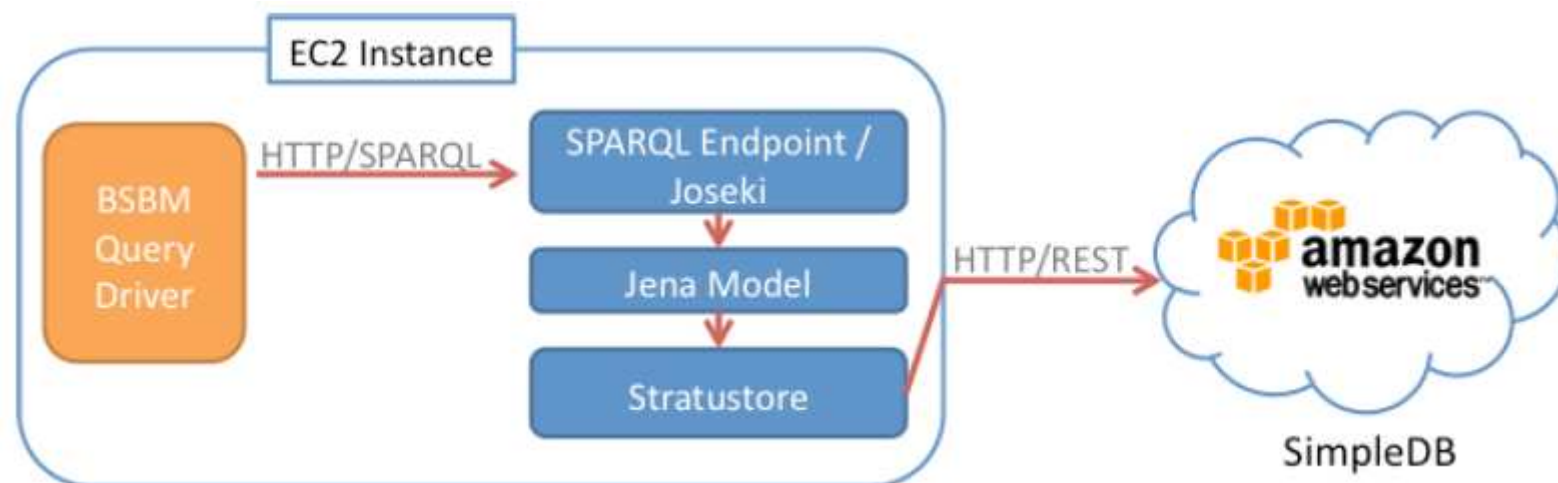
- Mismatch between RDF and SimpleDBs data organization
- Mismatch between SPARQL and SimpleDBs Select queries
- High-latency to access SimpleDB



```
SELECT s FROM products WHERE
  rdf:type = "example:Producer"
INTERSECTION
example:producesIn = "country:DE"
```

See NeFoRs paper / presentation
there for technical detail

- Berlin SPARQL Benchmark
 - Simulation data, queries and query driver for an (read heavy/ analytical) eCommerce use case (products, reviews etc)
- Evaluation Setup
 - Complete Setup is run on virtual machine on amazon EC2
 - Using 1,5,10 and 20 machines
 - 1 Million triples



Did it work?



- Is Stratustore the worry free triple store on the cheap?
- Up to a Point – the Positive:
 - Is easy to use, highly available, has automatic backup, pay per use
 - Created „on the cheap“ – just a jar that uses services out there
 - Scales elastically for many concurrent accesses for simple queries (can be faster than state of the art)
- The negative
 - Multiple orders of magnitude slower than state of the art for complex queries
 - Data structure restrictions also decrease generality of Stratustore
- The Potential
 - Many optimizations still possible
 - Without more complex query language (or at least API) for SimpleDB/DaaS a large performance gap will remain for complex queries

- Who?
- What?
- How?
- Does it work?
- And now?

Evaluating Services is Different



- Hard (and not desirable?) to control all variables
- Controlling and reporting the resources used is (sometimes) impossible
 - The service may react to the setting of the experiment, adapt to changing workload
 - This elasticity needs to be measured and reported
 - „Price“ takes the role of measuring resource usage
- -> We need a next generation of benchmarks

Sometimes the query is the answer



- Assumption: For many applications of RDF the majority of the queries follows a structure that a) is known in advance or b) that can be learned
 - -> We need approaches that utilize this knowledge for
 - targeted reasoning materialization
 - indexes
 - Are we too focussed on solutions that are schema & query unaware to utilize such information?



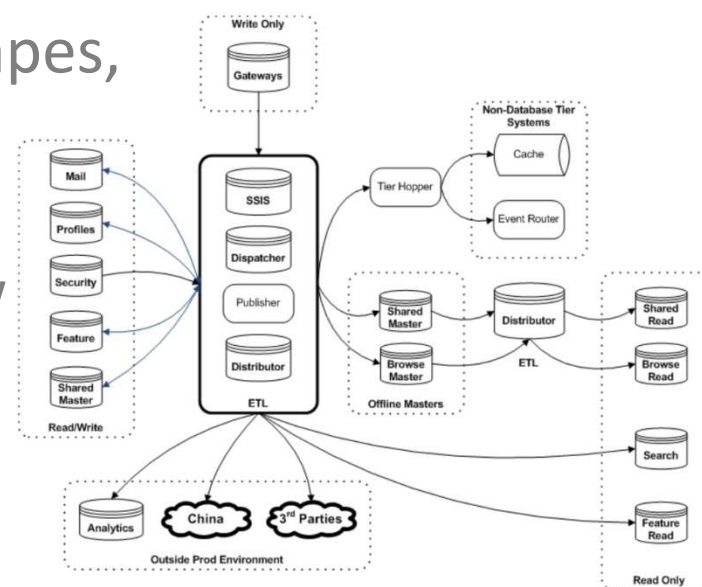
More Than Very Scalable Reasoning



- Cloud Computing & Semantics is more than only distributed and scalable reasoning
 - **Simplicity:** Offering complex services as SaaS hides the complexity of running these services from the users
 - **Elasticity:** Services that can adapt to shifting requirements, in particular to sudden load spikes
 - **Utility Pricing:** Services that can charge in a transparent way for a user's resource consumption
 - **Service Orchestration:** Not software developed from ground up, but value added on top of existing services

One Size Does Not Fit All – Not even RDF

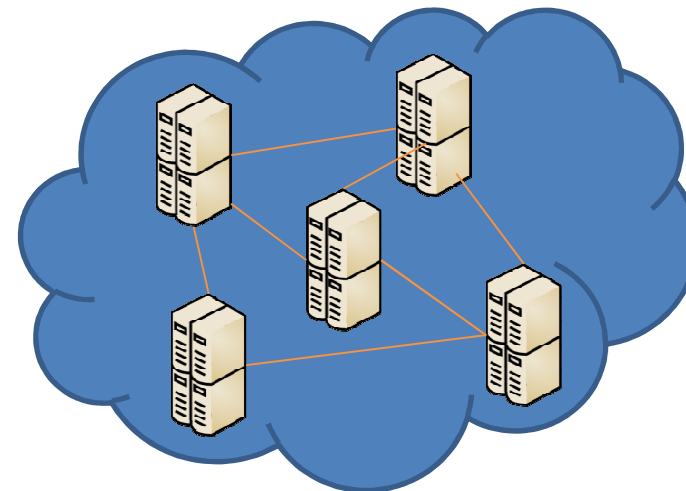
- Not all storage needs of current applications can be tackled with one RDBMS, increasingly we see storage landscapes build from many DBMS , with an increasing proportion of non-relational databases
 - -> RDF databases will be one specialized DB in these landscapes, only for the data they are well suited for
 - -> We have to think about how to manage these hybrid data storage landscapes



RDF on Cloud Number Nine

An Implementation, Experiment and Thoughts on Cloud Computing and Semantic Data Management

- Stratustor – the worry free triple store (up to a point)
- Evaluating services is different
- Sometimes the query the answer
- More than very scalable reasoning
- One size does not fit all



Thanks for your Attention!