Investigating Elastic Cloud Based Reasoning for the Semantic Web

Omer Dawelbeit, *MSc, MBCS*
School of Systems Engineering,
University of Reading

Introduction

- Part-time doctoral research at the School of Systems Engineering, University of Reading. Supervisor: Professor Rachel McCrindle.

- Full-time cloud computing professional at Appsbroker Consulting.

Background

- The current Web was designed for humans.
- Search is based on keyword occurrences.
- The Semantic Web [1] is an extension of the current Web that adds semantics to data.

WWW
Web of Documents

Web of Data

- Computers are able to answer queries based on the meaning of data.
- Semantic Web Data (knowledge) is represented in RDF or OWL.

The Semantic Web

- A statement of knowledge is called a Triple:

<http://domain/BCS> <rdf:type> <http://domain.com/#Organisation>

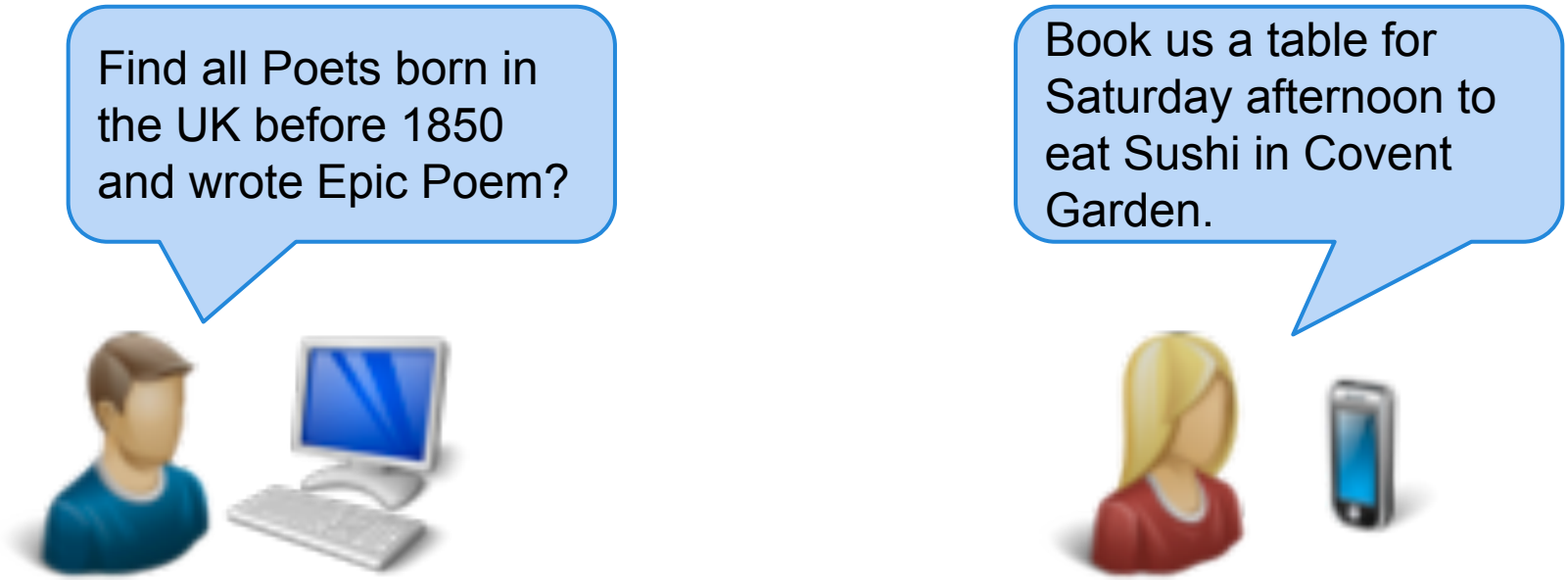
- Computers apply rule-based reasoning to infer new statements, for example:

Input: **<John> <rdf:type> <Manager>** and
<Manager> <rdfs:subClassOf> <Person>
Rule: if **a rdf:type B** and **B rdfs:subClassOf C** then **a rdf:type C**
Output: **<John> <rdf:type> <Person>**

- Computers infer knowledge using either forward reasoning or backward reasoning.

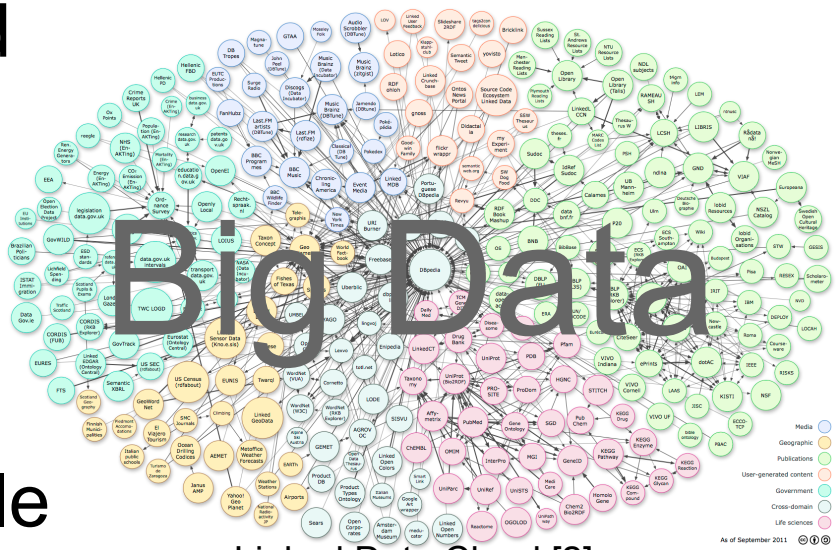
The Semantic Web

- The Semantic Web enables computers to:
  - Combine knowledge from different sources.
  - Infer new implicit knowledge.
  - Answer complex queries.



Problem Description

- The Semantic Web size is billions of statements.
- The data is highly skewed and inter-related.
- Reasoning and handling the data requires a great deal of computing power.
- Need efficient and scalable distributed/parallel algorithms.
- Need efficient data partitioning and assimilation between computing nodes.
- Need to cater for the storage of inferred knowledge.



Linked Data Cloud [2]

Aim and Objectives

- Review the state of large scale distributed Semantic Web reasoning.
- Investigate how cloud computing features (Elasticity and Big data services) can address the current issues.
- Identify factors impacting the cost of cloud based Semantic Web reasoning.
- Develop a framework for elastic, cost aware cloud based reasoning (ECARF).
- Evaluate the framework through a cloud based prototype.

Literature Review

- Embarrassingly Parallel [3] RDFS approach:
  - Shows linear scalability.
  - Generates large number of duplicates.
  - Hard to extend to support richer logic.
- MapReduce reasoning [4] on subset of OWL:
  - Very high throughput.
  - Difficult to store inferred data.
  - Difficult to extend to support richer logic.
- Peer to peer, Distributed Hash Tables [5,6]:
  - Loosely coupled commodity computers.
  - Suffers from load balancing issues.

Methodology

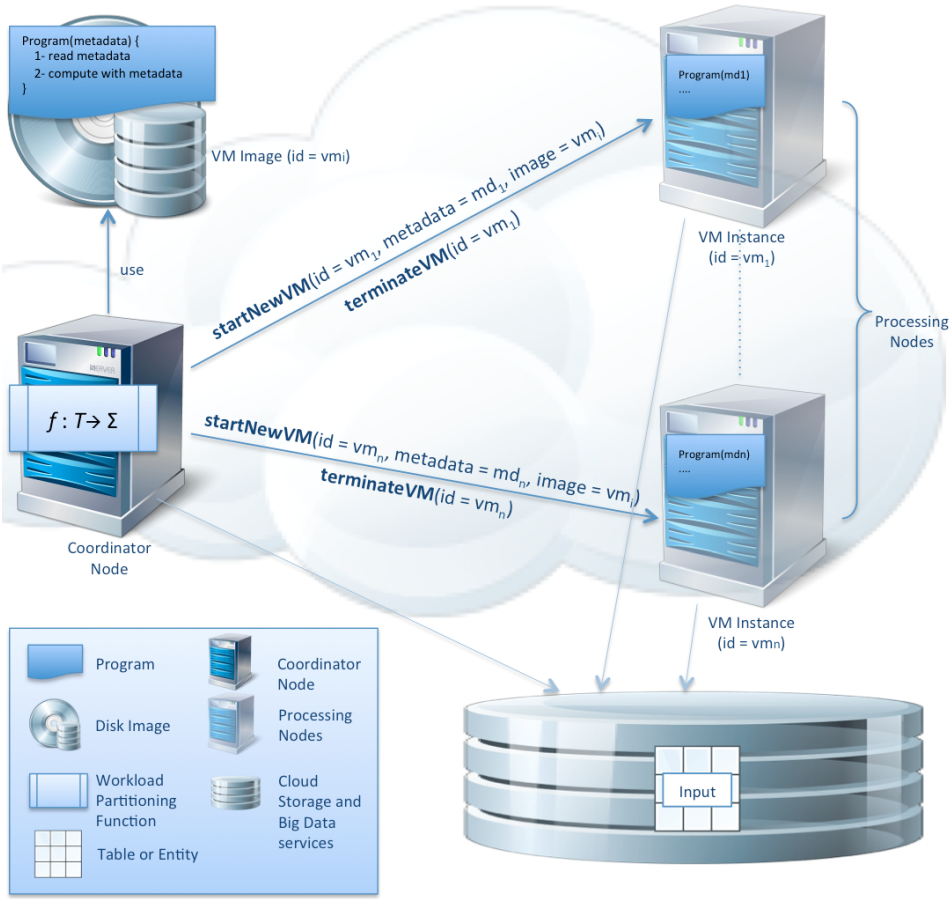
- Follow a design methodology.
- Literature review of current state of the art.
- Theoretical framework development (ECARF).
- Prototype development (AWS and/or GCP).
- Evaluation of ECARF in terms of cost, scalability and logic supported (RDFS, OWL 2 RL).
- Evaluation using a range of datasets:
  - DBpedia - Semantic extracts of Wikipedia.
  - SwetoDblp - Academic publications.
  - LUBM - Synthetic data benchmark.

Preliminary Framework

ECARF Formal Definition

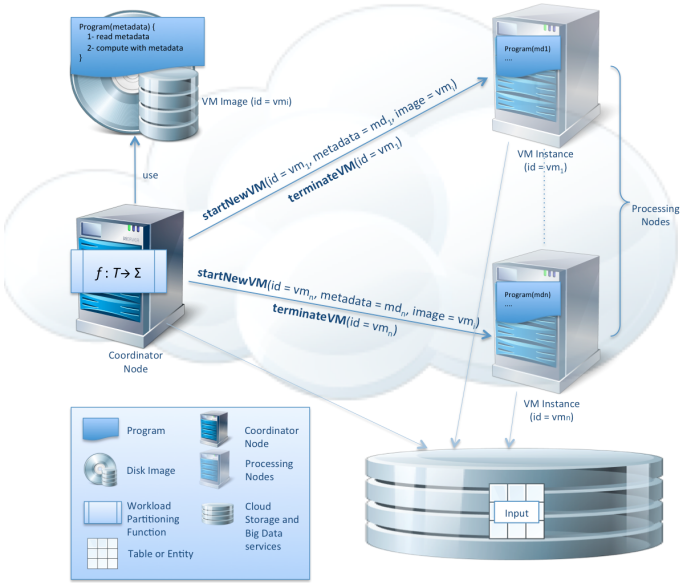
ECARF is a 8-tuple, $(\Theta, \Sigma, T, \Gamma, \Delta, \rho, f, \sigma)$, where $\Theta, \Sigma, T, \Gamma, \Delta$ are all finite sets and:

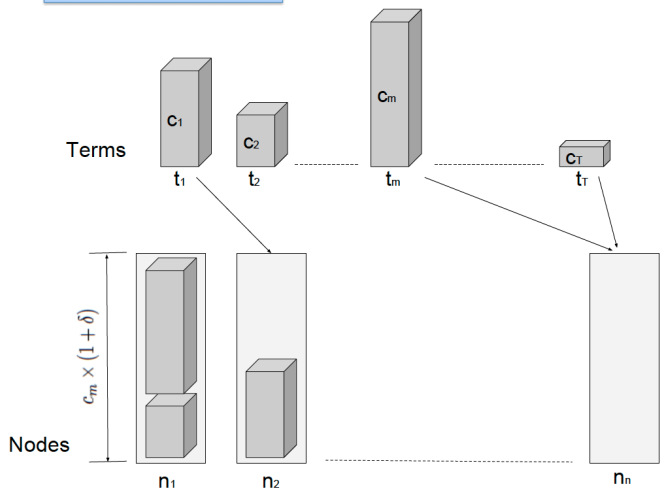
- Θ is the set of coordinators,
- Σ is the set of processing nodes,
- T is the set of work items,
- Γ is the set of input to be processed,
- Δ is the set of cloud services of mass cloud storage and big data service,
- ρ is the program embedded on a VM disk image and is capable of reading the node's metadata,
- $f : T \rightarrow \Sigma$, is the workload allocation function, and
- σ is the cost tracking function.



Preliminary Framework

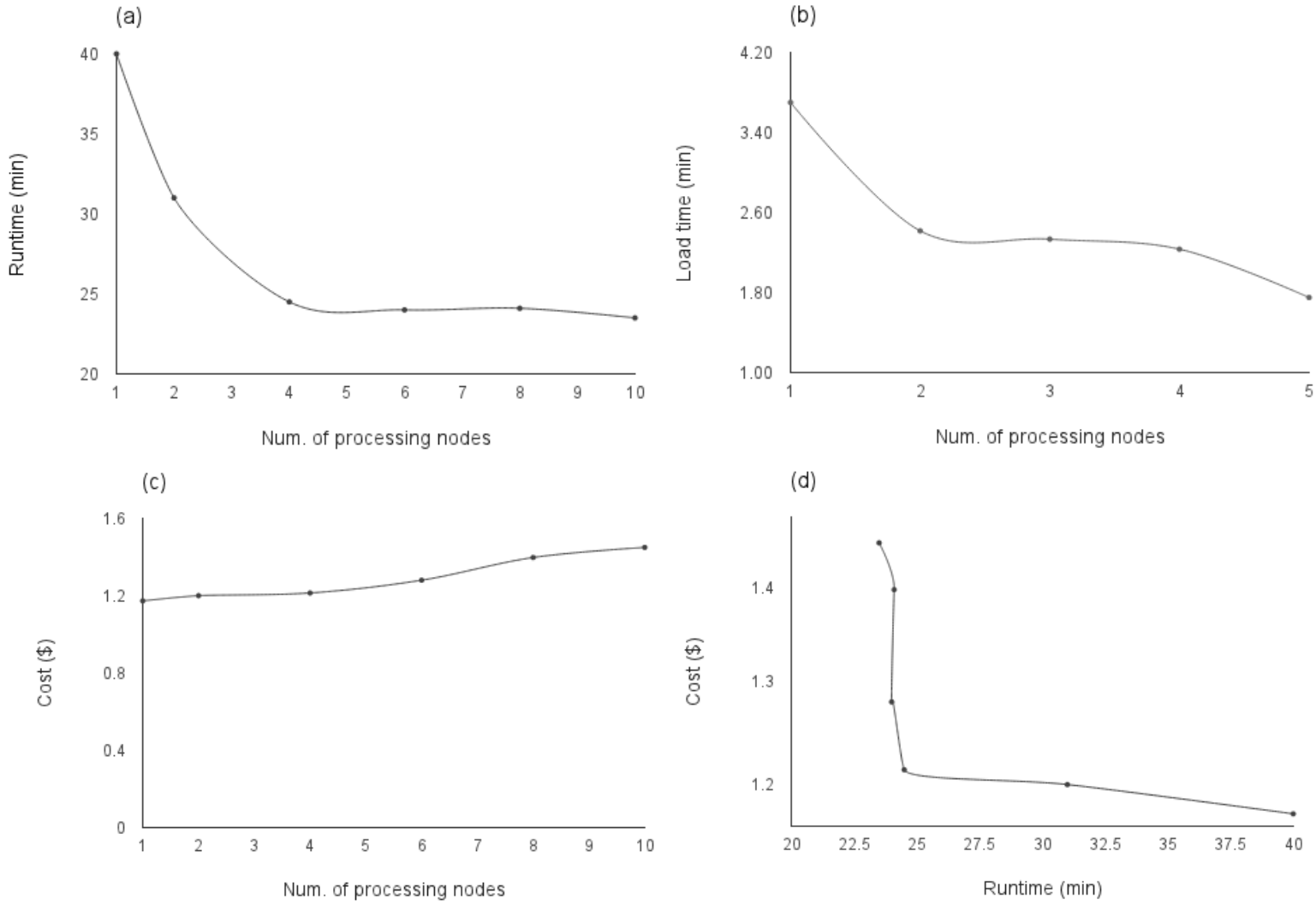
- Coordinator receives requests for work.
- Partitions the work using bin-packing algorithm [7].
- Starts the required number of nodes, supplying tasks as metadata.
- Monitors nodes and terminate them once done.
- The overall resource usage is tracked.





Bin Packing Algorithm

Initial Results



Conclusion

- Initial results are promising, closure of SwetoDblp in 24 minutes using only 4 nodes.
- Most of the processing of the massive data is handled by the cloud based big data services.
- Reported results outperforms DHT results.
- Results highlight the potential of the big data services for backward reasoning.
- To achieve Web-scale reasoning need to address the latency with data transfer.
- Need to also address issues with duplicate knowledge being inferred.

Thank you

Questions?

o.i.o.dawelbeit@
pgr.reading.ac.uk

http://omerio.com

@omerio

References

1. Berners-Lee, T., Hendler, J. and Lassila, O., 2001. The Semantic Web. Scientific American, 284(5), pp.34--43.
2. Linked Data Cloud, source (http://lod-cloud.net/versions/2011-09-19/lod-cloud_colored.html).
3. Weaver, J. and Hendler, J., 2009. Parallel materialization of the finite rdfs closure for hundreds of millions of triples. The Semantic Web-ISWC 2009, pp.682--697.
4. Urbani, J. et al., 2012. WebPIE: A Web-scale Parallel Inference Engine using MapReduce. Web Semantics: Science, Services and Agents on the World Wide Web, 10, pp.59--75.
5. Kaoudi, Z., Miliaraki, I. and Koubarakis, M., 2008. RDFS reasoning and query answering on top of DHTs. The Semantic Web-ISWC 2008.
6. Fang, Q. et al., 2008. Scalable distributed ontology reasoning using DHT-based partitioning. The Semantic Web, pp.91--105.
7. Coffman, E. G. Jr.; Garey, M. R.; and Johnson, D. S. "Approximation Algorithms for Bin-Packing--An Updated Survey." In Algorithm Design for Computer System Design. Vienna: Springer-Verlag, pp. 49-106, 1984