An e-Infrastructure enabled semantic search service

Nikos Simou & Costas Pardalis
National Technical University of Athens

Pilot Objectives

- Establish a search system using MICHAEL data
- Enrich the search system with semantic search capabilities
- Evaluate the feasibility of these requirements using e-infrastructures, presenting the main benefits from this integration

Cloud Platform

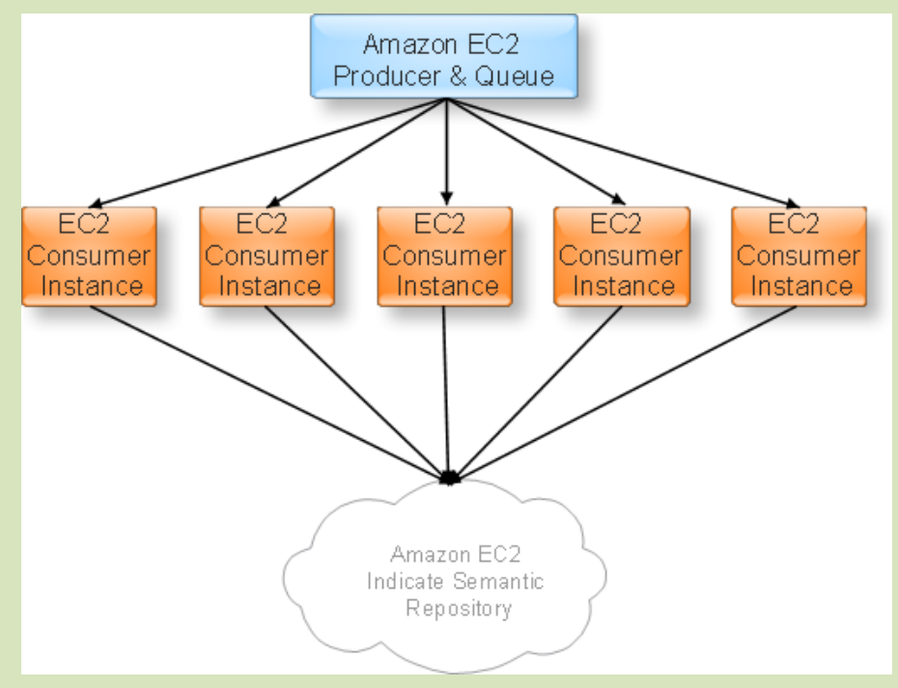
- Amazon EC2 is used as the Cloud environment for deployment.
  - It provides a concrete pricing model for comparisons.
  - It is one of the most technologically mature Cloud environments.

Amazon EC2 Utilized Services

- Amazon Elastic Compute Cloud
  - Large Instance 7.5 GB of memory, 4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each), 850 GB of local instance storage, 64-bit platform, were used to form the Indicate Cluster
- Elastic IP Addresses
  - Were assigned to each instance to ensure the existence of\ static IPs
- Amazon Elastic Block Store (EBS)
  - Was used for providing persistence storage to the Indicate Cluster Instances.

Data Manipulation @ Amazon EC2

- One Amazon EC2 Instance is acting as the producer and hosts the Message Queue (RabbitMQ).
- Five Large Amazon EC2 Instances are hosting the consumers.

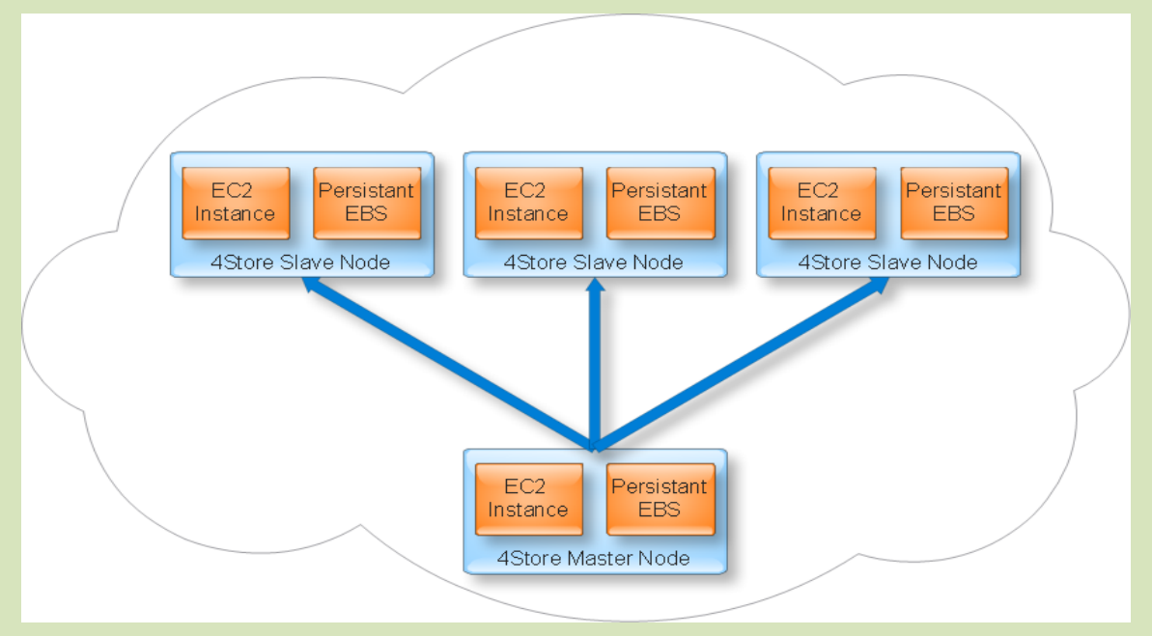


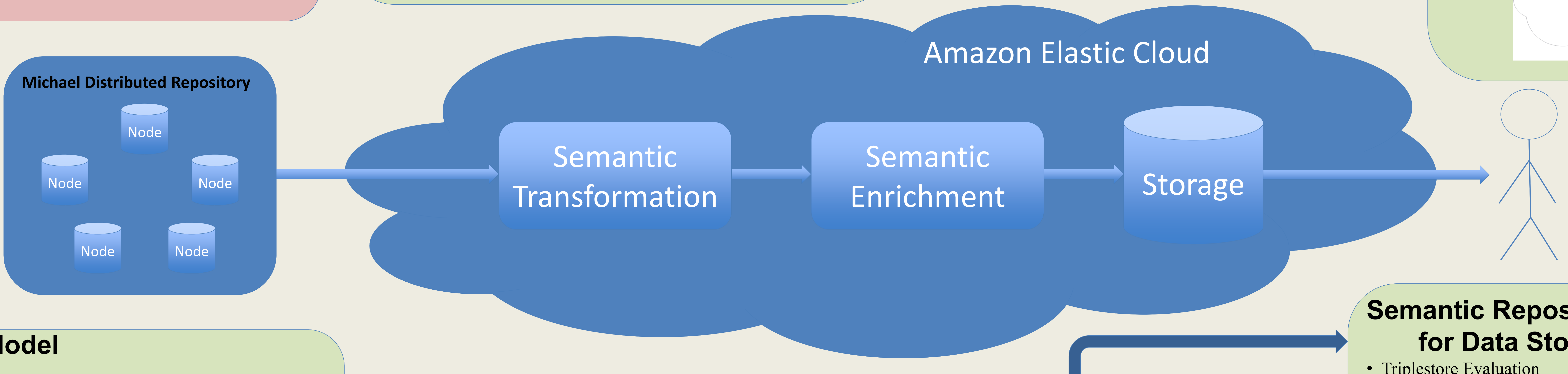
Evaluation

Method Used	Time in Millisecs
Local Host	22.383.937ms (~6.2hrs)
Local Cluster	5.020.430 (~1.39hrs)
Amazon Cloud	1.422.000 (~23.7 min)

Semantic Repository @ Amazon EC2

- The 4store Distributed Semantic Repository was installed on 4 Large EC2 Instances.
- The number of Nodes attached to the Semantic Repository can be adjusted in order to check scalability and performance.





Semantic Search

- Querying on data
  - Search for items from a specific country (e.g Greece)
- Semantic Querying
  - Search for items from a specific country (e.g Greece)
  - Search for items which are hold by Countries of Mediterranean Sea and are about alive politicians

Data Model

- Exploration of data
  - Every xml item represents a collection of digital cultural objects
- Mapping of xml elements to RDF properties for achieving semantic representation of data
  - Language → dcterms:language
  - Digital Format → dcterms:format

RDFization

XML Instance

```
<digital-collection id="UK-DC-
   2ee6a982">
 <identification>
     <title>Dambusters</title>
 </identification>
 <description>
     <digital-format-group>
           <en>JPEG</en>
      </digital-format-group>
 </description>
…
…

</digital-collection>
```

RDF Representation

```
<rdf:Description
   rdf:about="http://mint.image.ece.ntua.gr/re
   source/UK-DC-2ee6a982">
   <dc:title>Dambusters</dc:title>
   <dcterms:format>JPEG</dcterms:format>
   <dcterms:language>English<dcterms:language>
   <dc:subject>Defence</dc:subject>
   <dc:subject>Economic and social
   development</dc:subject>
   <dcterms:spatial>UNITED
   KINGDOM</dcterms:spatial>
        …
        …
</rdf:Description>
```

Semantic Enrichment

- Specific values of the examined dataset were discovered as DBpedia resources.
- Additional semantic information is added to the dataset
  - **Countries** : area, capital, density, currency, etc
  - **Languages** : spokenIn, languageFamily, speakers, etc
  - **Famous Persons** : dates of birth death, professions, works, etc

Semantic Repository for Data Storage

- Triplestore Evaluation
  - Requirements
    - a) Distributed
    - b) Licensing (open source)
    - c) Sparql language support
    - d) Web based access
- Candidates
  - 4store
  - Sesame
  - Bigowlim

Enrichment Results

	Total	Found	Percentage
Countries	16429	15987	97.3%
Languages	11090	11032	99.5%
Persons	6442	3632	56.4%

Conclusion

- Semantic Search using e-Infrastructures
  - Provides scalability that is vital for are semantic enrichment, since frequent updates required for remaining consistent.
- Cost
  - Processing: $ 0.68 per node per hour (~ 1.7 €)
  - Storage: $ 0.11 per Gb per month (~ 4.4 €)