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

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

Amazon Elastic Cloud

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

<table>
<thead>
<tr>
<th>Method Used</th>
<th>Time in Milisecs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Local Host</td>
<td>22.383.973/ms (~6.2hrs)</td>
</tr>
<tr>
<td>Local Cluster</td>
<td>5.020.430 (~1.39hrs)</td>
</tr>
<tr>
<td>Amazon Cloud</td>
<td>1.422.000 (~23.7 min)</td>
</tr>
</tbody>
</table>

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

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
- Languages → dcterms:language
- Digital Format → dcterms:format

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

RDFization
- XML Instance
- RDF Representation

Enrichment Results

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Found</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Countries</td>
<td>16429</td>
<td>15987</td>
<td>97.3%</td>
</tr>
<tr>
<td>Languages</td>
<td>11090</td>
<td>11032</td>
<td>99.5%</td>
</tr>
<tr>
<td>Persons</td>
<td>6442</td>
<td>3632</td>
<td>56.4%</td>
</tr>
</tbody>
</table>

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

Conclusion
- Semantic Search using e-Infrastructures
  - Provides scalability that is vital for an 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 €)