

EC2 AMI LinkedData

Idafen Santana Pérez
OEG-DIA
Facultad de Informática
Universidad Politécnica de Madrid
isantana@fi.upm.es

Nandana
Mihindukulasooriya
OEG-DIA
Facultad de Informática
Universidad Politécnica de Madrid
nmihindu@fi.upm.es

Boris Villazón-Terrazas
OEG-DIA
Facultad de Informática
Universidad Politécnica de Madrid
bvillazon@fi.upm.es

ABSTRACT

We present the process that has been followed for the development of an application that makes use of several heterogeneous datasets that are related to

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: Miscellaneous;
E.2 [Data storage representations]: Linked Representations

General Terms

Design, Experimentation

Keywords

linked data, linked open data

1. INTRODUCTION

Linked Data principles are being adopted by an increasing number of data providers, getting as a result a global data space on the Web containing billions of RDF triples [1].

Domains library, government, statistics, but no Computer Science related datasets why?

2. AMAZON EC2

Today the advancements of cloud computing has made it possible for us to have different environments for computing without having the need to have a dedicated infrastructure for each of the different environments we require. One of the main advantage of virtualized environment is that you can recreated the same environment as many times as you want with a minimal effort. This put the computer users lives at ease because they can use a virtual machines for different temporary tasks without having the trouble to change the configuration of the personal machines. This ability can be very useful in scenarios like reproducing environments for

software testing, to try out software before installing them on a physical machine i.e. as a staging machine.

If it is possible to use already prepared virtual machines which matches the requirements, for example, the architecture i.e. it is 32bit or 64bit, the operation system, storage and memory requirements it would save a lot of time and effort of people trying to prepare a computing environment for some given task. In other words, if we could facilitate the search and discovery of virtual machine images in a systematic way, we could build a lot of semantic application about this which will automate this process and make the use of cloud computing for these tasks much easier.

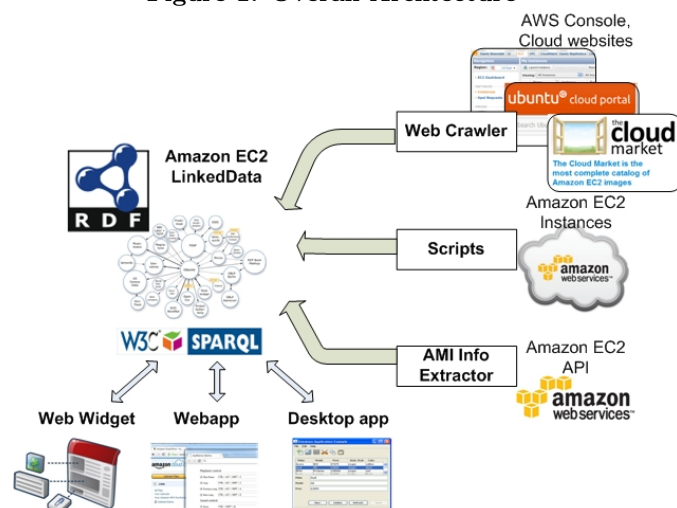
3. AMAZON EC2 LINKED DATA LIFE CYCLE

Here we briefly describe the proces

4. APPLICATIONS

Our goal of this project is to create a LinkedData dataset about Amazon virtual machine images. Thanks to this dataset, we have been able to build applications which can query the dataset using the defined ontology and help the users find Amazon Machine Images (AMI) that suit their requirements.

Figure 1: Overall Architecture



Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

I-SEMANTICS Triplification Challenge 2012 Graz, Austria
Copyright 2010 ACM 978-1-4503-0014-8/10/09 ...\$10.00.

4.1 Use cases for AmazonMachineImage reuse

Among many other advantages of cloud computing like scalability, high accessibility, and cost saving the ability to reuse a virtual machine as many times becomes a time saver in most of the cloud scenarios. Most cloud users create several master templates of the virtual machines images and use the appropriate one whenever necessary. This is far more efficient compared to setting up physical machines as the virtual machine images which are not currently executing will only consume storage space but no CPU power or memory. It is even better if one can use a public virtual machine image which satisfies the task at hand because it brings the setting up time to a bare minimum.

There are many use cases where it is very useful to find and reuse an existing virtual machine image without spending time on preparing the environment. For example, one might want to find a virtual machine environment which has a similar characteristics to a physical machine to experiment different configurations or settings without having the risk of corrupting the own machines. It also makes it possible to start things from the scratch with another instance of the same virtual image. Moreover, it is becoming common for software vendors to provide links to public virtual machine images which the users can use as a sandbox to first try the software on a virtual machine without installing on their own machines.

4.2 AmazonMachineImage data extraction

Amazon EC2 LinkedData dataset will be created by merging the data acquired by several different sources. The main source will be the API provided by AWS SDK for Java¹. An AMI Information Extractor java application will iterate through all the public Amazon Virtual Machine images and extract the data available from the API. However, the AWS SDK lacks the ability to extract several important attributes like operating system, or the image size which are essential to make this dataset useful in real life scenarios. The challenge is overcome by extracting information from other sources and aggregating them to the dataset. There are websites and portals which provide a richer set of attributes about Amazon Machine Images which contains the aforementioned information as unstructured data available as web pages. Examples of these sites include the AWS Management Console², Ubuntu Cloud Portal³, and The Cloud Market catalog <http://thecloudmarket.com/>. Needlebase platform⁴ is used to acquiring, integrating, and cleansing information from these sites and feed them to the Amazon EC2 LinkedData dataset. Even more detailed information about a given AMI can be extracted by a running script inside the Amazon instances created using the given AMI. Two scripts for Windows and Linux are made available for Amazon users which they can run to upload additional data about an AMI.

4.3 EC2LD applications

We have built three applications which consume the Amazon EC2 LinkedData and provide users AmazonMachineImage ids which match their requirements. These applications

are a web widget, a desktop application, a web application. The web widget which is implemented as a Google Gadget⁵ can be embedded in web pages. Once configured with certain properties, this widget can communicate with the SPARQL endpoint and dynamically fetch the Amazon virtual machine images which match the requirements. This can be a good addition to the software providers i.e. they can include this widget in the download pages so that users know in which public virtual machine images they can try this software.

The VMI Finder desktop application can run on a computing environment and gather the data and then will create a SPARQL query based on those information to query the SPARQL endpoint exposed by Amazon EC2 LinkedData. Users can manually modify the information that has been automatically collected before executing the query by editing, adding, or removing any information. The query result will provide a set of AmazonMachineImage ids which matches the computing environment it runs. The web application also provides a similar functionality where users can directly search for AmazonMachineImages via a web interface.

5. CONCLUSIONS AND FUTURE WORK

In this paper we have presented an application that makes use of several ...

6. ACKNOWLEDGMENTS

This work has been supported by the R&D project Seals. We would like to kindly thank

7. REFERENCES

¹<http://aws.amazon.com/documentation/sdkforjava/>

²<http://aws.amazon.com/console/>

³<http://cloud.ubuntu.com/>

⁴<http://needlebase.com/>

⁵<https://developers.google.com/gadgets/>