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

In recent years we have seen the two most substantial transformations in the ICT world, that are virtualization and cloud computing. Virtualization helps enterprises make more efficient use of hardware resources. It facilitates a greater degree of abstraction of the software environment from its hardware. Servers now exist as a single file, as a virtual machine. It is possible to easily move them from one piece of hardware to another, duplicate them at will, and create a more scalable and flexible infrastructure. However the question of how virtual machines data can be integrated into the Web of Data has not yet been sufficiently addressed. In this paper, we present the process that has been followed for the development of Linked Data applications that facilitate the search and discovery of virtual machine images in a systematic way.

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, virtual machine

1. INTRODUCTION

Virtualization is a computing technology that enables a single user to access multiple physical devices. Virtualization may also be used for running multiple applications on each server rather than just one; this in turn reduces the number of servers companies need to purchase and manage. It enables to consolidate servers and do more with less hardware. In the other hand, cloud computing offers scalable infrastructure and software off site, saving labor, hardware, and power costs. Financially, the cloudâĂŹs virtual resources are typically cheaper than dedicated physical resources connected to a personal computer or network. With cloud computing, the software programs are not running from the personal computer, but rather are stored on servers housed elsewhere and accessed via the Internet. [2].

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]. Moreover, Linked Data technologies are being using to share data covering a wide range of different topical domains, such as Media, Geographic, Government, Publications, Cross-domain, and Life sciences. However, there are no datasets that cover the Computer Science domain, which means that we, as computer scientists, are not taking the advantages of adopting Linked Data Principles to our own domain.

This paper aims at showing that is possible to adopt Linked Data Principles within the Computer Science domain, specifically in the virtualization and cloud computing fields. In this paper, we present the process that has been followed for the development of Linked Data applications that facilitate the search and discovery of virtual machine images in a systematic way. The rest of paper is organized as follows: Section 2 introduces the Amazon Elastic Compute Cloud (EC2) AMI, Section 3 explains the process we followed for the generation of linked data, Section 4 describes the linked data applications that we built, and finally, Section 5 presents the conclusions and future work.

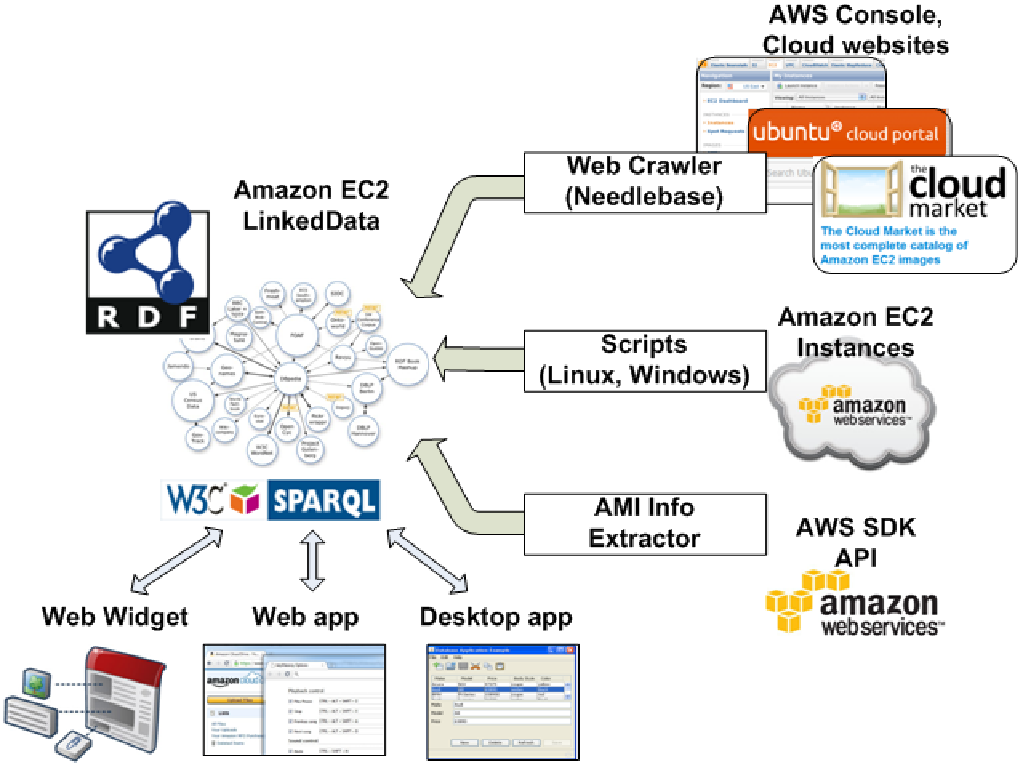
2. AMAZON EC2 AMI

Nowadays the advancements of cloud computing have 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 it is possible to recreate 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 32 bits or 64 bits, the operation system, stor-

I-SEMANTICS Triplification Challenge 2012 Graz, Austria

Figure 1: High Level Architecture



age and memory requirements it would save a lot of time and effort of people trying to prepare a computing environment for a 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 applications on top of this that will automate this process and make the use of cloud computing for these tasks much easier.

Amazon Elastic Compute Cloud (EC2) is a central part of Amazon.com's cloud computing platform, Amazon Web Services[1] (AWS). EC2 allows users to rent virtual computers on which to run their own computer applications. EC2 allows scalable deployment of applications by providing a Web service through which a user can boot an Amazon Machine Image to create a virtual machine, which Amazon calls an "instance", containing any software desired. A user can create, launch, and terminate server instances as needed, paying by the hour for active servers, hence the term "elastic". EC2 provides users with control over the geographical location of instances that allows for latency optimization and high levels of redundancy.

3. EC2 AMI LINKED DATA LIFE CYCLE

In this section we briefly describe our process for generating, interconnecting, and publishing EC2 AMI linked data. This process was inspired by existing methodological guidelines [3], which propose an iterative incremental life cycle model where EC2 AMI LD gets continuously improved and extended. The EC2 AMI Linked Data life cycle consists of five main activities, namely specification, modelling, generation, publication and exploitation. We focus on the description of the (1) specification of the resultant application, section 3.1, and (2) data extraction and RDF transformation, sectionv3.2.

The goal is to create a LinkedData dataset about Amazon virtual machine images. Thanks to this dataset, we are able to build applications that can query the dataset using the defined ontology and help the users find Amazon Machine Images (AMI) that suit their requirements. Figure 1 shows the high level architecture.

3.1 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 bring 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 make it possible to start things from the scratch with another instance of the same virtual image. Moreover, it is becoming common for

[1] http://aws.amazon.com/ec2/

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.

3.2 Amazon Machine Image data extraction

Amazon EC2 LinkedData dataset is created by merging the data acquired by several different sources. The main source is the API provided by AWS SDK for Java [2]. 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 provides a richer set of attributes about Amazon Machine Images which contains the aforementioned information as unstructured data available as wep pages. Examples of these sites include the AWS Management Console [3], Ubuntu Cloud Portal [4], and The Cloud Market catalog http://thecloudmarket.com/. Needlebase platform [5] 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 scripts 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.

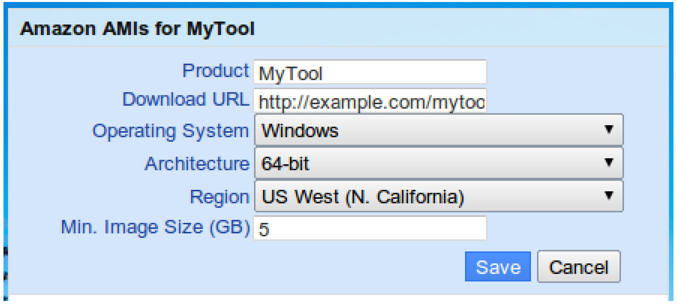
Finally, it is worth mentioning that the SPARL endpoint is available at `http://mccarthy.dia.fi.upm.es:8892/sparql`.

4. EC2LD APPLICATIONS

We have built three applications that 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 Gadget [6] 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. Figure 2 shows an example of the parameters. 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 provide a similar functionality where users can directly search for AmazonMachineImages via a web interface. Figure 3 depicts the search results.

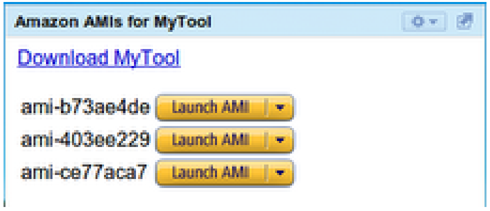
Figure 2: Configuration parameters of the web widget



Amazon AMIs for MyTool

Product: MyTool
Download URL: http://example.com/mytoo
Operating System: Windows
Architecture: 64-bit
Region: US West (N. California)
Min. Image Size (GB): 5

Figure 3: Search results of the web widget



Amazon AMIs for MyTool

Download MyTool

ami-b73ae4de
ami-403ee229
ami-ce77aca7

The applications are available at `http://mccarthy.dia.fi.upm.es/ec2ld/`

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have presented the process that has been followed for the development of Linked Data applications that facilitate the search and discovery of virtual machine images in a systematic way. As future work we will include more virtual machines from other vendors.

6. ACKNOWLEDGMENTS

This work has been supported by the R&D project Seals (FP7-238975).

7. REFERENCES

[1] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*, volume 1. Morgan Claypool, 2011.

[2] M. Steinder, I. Whalley, and D. Chess. Server virtualization in autonomic management of heterogeneous workloads. *SIGOPS Oper. Syst. Rev.*, 42(1):94–95, Jan. 2008.

[3] B. Villazón-Terrazas, L. Vilches-Blázquez, O. Corcho, and A. Gómez-Pérez. Methodological guidelines for publishing government linked data. In D. Wood, editor, *Linking Government Data*, pages 27–49. Springer, 2011.

[2] `http://aws.amazon.com/documentation/sdkforjava/`

[3] http://aws.amazon.com/console/

[4] http://cloud.ubuntu.com/

[5] http://needlebase.com/

[6] https://developers.google.com/gadgets/