DOI: http://dx.doi.org/10.5769/C2011024

"Forensic as a Service - FaaS"

Dener Didoné[a], Ruy J. G. B. de Queiroz[b]
"(a) Universidade Federal de Pernambuco - Centro de Informática – dd@cin.ufpe.br, Recife, Brazil"
"(b) Universidade Federal de Pernambuco - Centro de Informática – ruy@cin.ufpe.br, Recife, Brazil"

Abstract — The extent to which cloud computing has become ubiquitous, we have experienced big changes in paradigms, from the time of centralized computing (mainframes) through decentralization and re-centralization interconnected via the Internet. The definition of what is cloud still causes great confusion in the industry, but the cloud is an evolution of the Internet that allows everything to be delivered as a service - Everything as a Service (EaaS, XaaS, *aaS). Our purpose is to demonstrate the use of the cloud to provide Forensic as a Service (FaaS) through flexible, elastic and dynamic platforms such as storage and processing power to "unlimited", besides demonstrating that the use of Forensic as a Service becomes an interesting alternative when working on large data sets.

Keywords — cloud forensics; distributed forensic; forensic as a service; cloud computing; forensic computing

1. Introduction

The ubiquity offered by the Internet plus the low cost of electronic equipment such as computers, notebooks, PDAs, tablets and more recently the progress of the Smartphone, has made computer crimes become frequent. The effort made by law enforcement and governments to solve these cases is immense, requiring investments of millions of dollars in training and infrastructure, and upgrading the legislative system.

The computer forensic is the support for solving these crimes, this being a relatively new discipline, an evolution of techniques for data recovery, which has collaborated in obtaining evidence for the elucidation of the crimes.

Cloud computing has become a trendy nowadays, providing "unlimited" computational resources (such as storage, network and processing) for organizations and governments at low cost, offering scalable, on-demand and pay-per-use model, without worrying with infrastructure, configuration and management of hardware issues.

With the increasing demand of electronic crimes added by a factor cited by Garfinkel (2010) as one of the problems in view of computer forensic, which is "the growing size of storage devices means that there is frequently insufficient time to create a forensic image of a subject device, or to process all of the data once it is found.".

As a way to collaborate in solving the capacity increase of storage devices, this article aims to present the proposal for delivery Forensic as a Service - FaaS, using the concept of cloud computing and distributed computing with the programming model MapReduce.

Will be briefly discussed the concepts of computer forensic, cloud computing, MapReduce delivery of computing "... as a Service" model. We will discuss security and some legal issues, addressing evaluation and validation of the proposal.

2. Background

2.1. Forensic computing

Forensic Computing has approximately 40 years old, and what we actually know as forensic techniques were developed primarily for data recovery. "Within the past few years a new class of crime scenes has become more prevalent, that is, crimes committed within electronic or digital domains, particularly within cyberspace. Criminal justice agencies throughout the world are being confronted with an increased need to investigate crimes perpetrated partially or entirely over the Internet or other electronic media. Resources and procedures are needed to effectively search for, locate, and preserve all types of electronic evidence. This evidence ranges from images of child pornography to encrypted data used to further a variety of criminal activities". (Lee, Palmbach & Miller 2001).

Since then, forensics in electronic devices (PDA, phones, tablets, notebooks, etc.) has become common and necessary for elucidation of electronic crimes. According to the FBI in 2010 were examined more than 3.300 TB of data processed, and an analysis effort that yielded more than 12 TB data/day (considering only working days). (Regional Computer Forensics Laboratory [RCFL], 2010).

The computer forensic is a multidisciplinary science that applies investigative techniques to determine and analyze evidences and in some cases to form a hypothesis as to their cause and effect. It follows the methods and procedures defined for the four forensics phases: identification, analysis, preservation and presentation of data.

There is a need to take reasonable care that the evidence does not change during the course of forensic, this is a constant concern, because once started any type of system or file, and it undergoes changes in its content or in its meta-content. Starting up an operating system, for example, generates a cascading of events that trigger the generation of logs provided by configuration changes and consequently changes in the evidence.

From the exposed, we realize the importance of computer forensics in the current scenario, allowing the identification of the authors of crimes, discovering clues through network data, disk images, intrusion logs generated by the operating system, among others, in short, Computer Forensic is an important tool to aid in maintaining the well being of civilized society, with respect to compliance with the established laws.

2.2. Cloud computing

Even more recent and more controversial, cloud computing brings a new configuration in the use of known technologies, being a generic term that can be defined as the evolution of technologies and processes, compound services, applications, distributed information and infrastructure, so that these can be arranged in a dynamic, resilient and quick way, to the extent that they are consumed. (Marins, 2009)

The cloud offers on-demand services, like electricity or gas, for example, being a natural evolution of the Internet and virtualization technologies, service oriented architecture and utility computing. The details are abstracted from the end-user who doesn't need expertise in the field of infrastructure of the cloud to use it.

The Cloud concept, according Vaquero, Rodero-Merino, Caceres and Lindner (2009) is still changing and the definition is conceived today. We adopted definition given by the National Institute of Standards and Technology – NIST (Mell & Grance, 2009), cloud computing is a model that allows convenient access and on-demand using the network, to a series of shared computing resources and configurable, where these resources can be quickly and easily provisioned or released with minimal management effort or service provider interaction.

We can found a minimum common denominator in the Cloud Definition that is scalability, pay-per-use utility model and virtualization. As is highlighted by Buyya, Yeo and Venugopal (2008) all of these computing services need to be highly reliable, scalable, and autonomic to support ubiquitous access, dynamic discovery and composability. In particular, consumers can determine the required service level through Quality of Service (QoS) parameters and Service Level Agreements (SLAs).

NIST (2009) has a vision that defines the cloud computing model in five essential characteristics, three service models and four deployment models. These characteristics are already widely known and disseminated and are specified below:

- On-demand self-service: A consumer can unilaterally provision computing capabilities such as server time and network storage as needed automatically, without requiring human interaction with a service provider.
- Broad network access: Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and PDAs).
- Resource pooling: The provider's computing resources are pooled to serve multiple consumers using a multi-tenant model.
- Rapid elasticity: Capabilities can be rapidly and elastically provisioned to quickly scale out; and rapidly released to quickly scale in.
- Measured service: Cloud systems automatically control and optimize resource usage by leveraging a metering capability at some level of abstraction appropriate to the type of service .

Cloud services offered by the cloud model can be classified in three distinct models that we have defined below according NIST (2009) and Vaquero et al. (2009):

- Software as a Service – SaaS: an alternative to locally run applications, running on a cloud infrastructure;
- Platform as a Service – PaaS: the capability to deploy onto the cloud infrastructure consumer-created or acquired applications created using programming languages and tools (supported by the provider).
- Infrastructure as a Service – IaaS: through virtualization, we can split, assign and dynamically resize storage, networking and processing capacity, as demanded by customers, providing fundamental computing resources.

This four deployment models are defined below by NIST (2009):

- Private cloud: The cloud infrastructure is operated solely for an organization. It may be managed by the organization or a third party and may exist on premise or off premise.
- Community cloud: The cloud infrastructure is shared by several organizations and supports a specific community that has shared concerns.
- Public cloud: The cloud infrastructure is made available to the general public or a large industry group and is owned by an organization selling cloud services.
- Hybrid cloud: The cloud infrastructure is a composition of two or more clouds (private, community, or public) that remain unique entities but are bound together by standardized or proprietary technology that enables data and application portability.

Taking into the account the characteristics, service and deployment models offered by cloud computing emerges a recent discussion about Everything as a Service – *aaS, nomenclature adopted to any kind of computation that can be provided as a service, and we include the possibility of provide Forensic as a Service – FaaS, what is a new opportunity to provide computing resources for forensic computing, including all intrinsic values embedded in the cloud, which include "unlimited" processing and storage in your usage on-demand self service and elasticity.

2.3. *Distributed forensic*

The use of distributed computing for conduct of forensic tasks is presented in this proposal through the implementation of the MapReduce programming model on the cloud. In our proposal we used Hadoop that is a software framework developed as an open-source implementation of the MapReduce programming model, supported by Apache. In the subsections below we have their definitions.

2.3.1. *MapReduce*

MapReduce is a distributed programming paradigm, developed by Google to simplify the development of scalable, massively-parallel application that process Terabytes of data on large commodity clusters.

Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system. (Dean & Ghemawat, 2004)

Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model, including forensics tasks (string operations, image processing, statistical analysis, etc). (Roussev, Wang, Richard III & Marziale, 2009)

The canonical example of MapReduce is a process to count the appearances of each different word in a set of documents, presented at figure 1.

In this example, retrieved from Dean & Ghemawat (2004), each document is split into words, and each word is counted initially with a "1" value by the Map function, using the word as the result key. The framework puts together all the pairs with the same key and feeds them to the same call to Reduce, thus this function just needs to sum all of its input values to find the total appearances of that word.

2.3.2. *Hadoop*

Hadoop was developed as an open-source implementation of the MapReduce programming model. This is a top-level Apache project being built and used by a global community of contributors, using the Java programming language.

Your architecture is composed by Hadoop Common, which contains the files needed to run Hadoop and support for the Hadoop's subprojects - Hadoop Distributed File System (HDFS) and MapReduce.

According with Apache Hadoop - HDFS Architecture Guide (2010), the HDFS is a distributed file system designed to run on commodity hardware, being highly fault-tolerant and designed to be deployed on low-cost hardware, providing high throughput access to application data and is suitable for applications that have large data sets.

Consulted Apache Hadoop – MapReduce Tutorial (2010) says that a MapReduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks.

The MapReduce framework consists of a single master JobTracker and one slave TaskTracker per cluster-node. The master is responsible for scheduling the jobs' component tasks on the slaves, monitoring them and re-executing the failed tasks. The slaves execute the tasks as directed by the master.

Minimally, applications specify the input/output locations and supply map and reduce functions via implementations of appropriate interfaces and/or abstract-classes. These, and other job parameters, comprise the job configuration. The Hadoop job client then submits the job (jar/executable etc.) and configuration to the JobTracker which then assumes the responsibility of distributing the software/configuration to the slaves, scheduling tasks and monitoring them, providing status and diagnostic information to the job-client.

Although the Hadoop framework is implemented in Java, MapReduce applications need not be written in Java. Hadoop Streaming is a utility which allows users to create and run jobs with any executables (e.g. shell utilities) as the mapper and/or the reduce.

3. Proposal

Anticipating the crisis in computer forensic, Garfinkel (2010) cites the growth in capacity of storage devices as one of challenging topics, adding that often not enough time

```
void map(String name, String document):
  // name: document name
  // document: document contents
  for each word w in document:
    EmitIntermediate(w, "1");
```

```
void reduce(String word, Iterator partialCounts):
  // word: a word
  // partialCounts: a list of aggregated partial counts
  int result = 0;
  for each pc in partialCounts:
    result += ParseInt(pc);
  Emit(AsString(result));
```

Fig. 1. Canonical example of MapReduce.

to create a forensic image of a disk, or process all the data found.

A brainstorming session at Computer, Information, and Systems Sciences, and Engineering (CISSE) 2008 explored researches categories, topics and problems in digital forensics. One of the results of this was an article where Nance, Hay and Bishop (2009) identified six categories for digital forensic research, between them a category called Data Volume where the authors says that parallelization of data processing could provides benefits to the area.

As previously mentioned, the quantity of devices coupled with the growth of storage space and the cheapening of technology makes terabytes of data to be analyzed in cases of cybercrime. The problem is that it ends by extrapolating the computational power of single workstations used in forensics laboratories by experts today.

An interesting approach utilizes clusters and large-scale distributed resources has demonstrated efficiently in typical forensic functions, how in Roussev and Richard III (2004), where the great potential of using cluster resources are demonstrated an early prototype and experiments, that speed up typical forensic functions. The authors also emphasize the need for forensic analysis techniques more sophisticated, which may be enabled by the transition of current tools for distributed tools.

An effort was done by Tang and Daniels (2005) that created a simple framework for network forensics based on the distributed techniques thereby providing an integrated platform for automatic forensics evidence collection and storage, including a mechanism for generating attack graphs to illustrate hacking procedures. Another research effort is the ForNet - Shanmugasundaram, Memon, Savant and Bronnimann (2009), aimed at distributed forensics, which is mostly concerned with the distributed collection and query of network evidence.

A recent proposal presented by Roussev et al. (2009) describe the efforts aiming at a new and open implementation of the MapReduce processing model, which significantly outperforms prior work on typical forensic tasks in large datasets, demonstrating linear scaling for CPU-intensive processing and even super-linear scaling for indexing-related workloads.

Roussev et al. (2009) uses an approach to accommodate more processing in the same amount of time, supporting the utilization of more hardware resources, which allows more machine resources to be deployed. The focus of Roussev is exclusively on supporting the use of commodity distributed computational resources to improve the turnaround time of forensic investigations.

Besides these studies, commercial tools such as Forensics Toolkit (FTK) from AccessData, are beginning to show interest and effectiveness in the use of forensic distributed, including support for multi-core processors, and distribution of processing, but with still limited distribution (limited to four machines).

Once mentioned related work, beyond the essential features and benefits brought by the cloud, demonstrating and defining viable features that allow the implementation of the MapReduce we try to present our idea. Our model aims at demonstrating the use of the cloud to provide Forensic as a Service (FaaS) through Cloud Computing, demonstrating flexible, elastic and dynamic platforms with "unlimited" storage and processing power, together with the MapReduce computation model in your open-source implementation Hadoop, besides demonstrating that the use of Forensic as a Service becomes essential, interesting and affordable for businesses, government or individuals in need.

Although the MapReduce programming model does not answer all the problems of distribution of processing, so far has demonstrated great ability in solving problems related to forensic tasks, such as string operations, image processing, statistical analysis, etc., that will be demonstrated later.

Figure 2 illustrates the scenario of our proposal, where users can use any devices to access servers through the

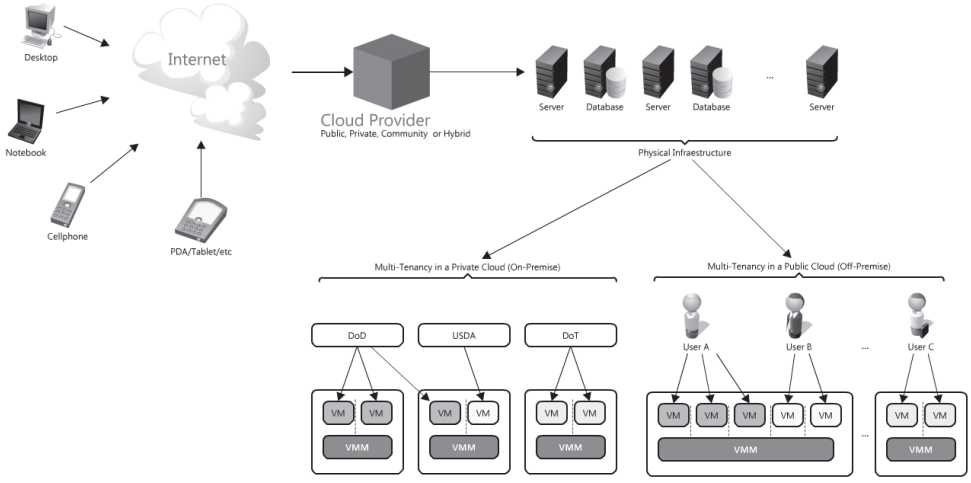

Fig. 2. Proposed model for delivery of Forensic as a Service (FaaS)

cloud and then perform forensics tasks, taking advantage of functions implemented by Hadoop MapReduce.

In the figure are presented the multi-tenancy model in private cloud, taking advantage of on-premise, and in public clouds, taking advantage of distributed costs among users and off-premise model.

We will present two options for implementing the proposed model, both in public and private clouds in the models on-premise and off-premise. The choice between public and private clouds differentiates itself in the issue of information security, security of the cloud and using its own technological park, which will be argued in your respective session.

3.1. Implementations on Public and Private Cloud

Our choice of public implementation will be through services offered by Amazon Web Services (AWS), the Amazon Elastic Map Reduce, a web service that enables businesses, researchers, data analysts, and developers to easily and cost-effectively process vast amounts of data. It utilizes a hosted Hadoop framework running on the web-scale infrastructure of Amazon Elastic Compute Cloud (Amazon EC2) and Amazon Simple Storage Service (Amazon S3). (Amazon Web Services [AWS], 2011)

Public clouds like Amazon derives a robust structure, easily manageable and with many options for scalability, being one of the most widely known and trusted provider of Cloud services, but raising issues of security, privacy and legal issues not yet resolved, which will be discussed later.

Utilization of this service is simple, you use the AWS Management Console. In this console you have full control of Jobs being executed or finalized. You just need these simple steps to use the service: Develop your data processing application (SQL-like, Pig, Hive, Cascading, Java, Ruby, Perl, Python, PHP, R, or C + +); Upload your data and your processing application into storage offered by Amazon, Amazon S3 (you use it as input data to be analyzed and output its results); Log into control panel offered by Amazon (AWS Management Console) and start your "Job Flow, "setting the types and quantities of Amazon EC2 instances to be used (there is also the option of using command line or APIs); You can monitor the process (when finished you can check the results on Amazon S3); When its work is finalized the Amazon EC2 instances are switched off (you only pay for what you actually consumed).

Some highlights of this service are covered by the definition of the essential characteristics of cloud computing, we quote them briefly: flexibility in the use of bodies as required, ease of use, because you don't need to worry about setting up, running, or tuning the performance of Hadoop clusters; integration with another Amazon's systems (S3 e EC2); inexpensive due the benefit of Amazon's scale, that you need to pay a very low rate for the compute capacity that you actually consume; multiple locations that you can choose the most appropriate; and finally, the use of third part tools like Karmasphere Studio Community Edition, that supports Amazon Elastic MapReduce and offer a free graphical environment to develop, debug, deploy and monitor MapReduce jobs from your desktop directly to Amazon Elastic MapReduce.

The payment is made for service used, in the pay-per-use model paying only for what you really consumed. All prices of this services can be found in pages of the respective products, Elastic Compute Cloud - EC2 (htpp://aws.amazon.com/ec2), Simple Storage Service - S3 (http://aws.amazon.com/s3) and Elastic MapReduce (http://aws.amazon.com/elasticmapreduce).

After presenting a deployment solution in public clouds, we cite the use of private clouds, which eventually supplement or even mitigate security issues still fresh in the use of cloud, which will be discussed later.

Our solution for deploying private clouds is the Eucalyptus (Elastic Utility Computing Architecture Linking Your Programs To Useful Systems), that is an open source software infrastructure for implementing on-premise clouds on existing Enterprise IT and service provider infrastructure. Thus, with a Eucalyptus private cloud, sensitive data remains secure from external intrusion behind the enterprise firewall.

Eucalyptus has design principles that ensure compatibility with existing Linux-based data center installations, can be deployed without modification on all major Linux OS distributions (Ubuntu, RHEL/CentOS, openSUSE, and Debian). The software framework is a highly modular cooperative set of web services that interoperate using standard communication protocols. Through this framework it implements virtualized machine and storage resources, interconnected by an isolated layer-2 network. (Nurmi et al., 2008) (Eucalyptus System, 2009)

The conceptual representation of the Eucalyptus Cloud is presented in Eucalyptus System (2009) that specifies: CLC is the Cloud Controller which virtualizes the underlying resources (servers, storage, and network). The Cluster Controllers (CCs) form the front-end for each cluster defined in the cloud. NCs are the machines on which virtual machine instances run. The Storage Controller (SC) provides block storage service (similar to Amazon EBS) while the Walrus storage system spans the entire cloud and is similar to the Amazon S3 in functionality. A Management Platform provides a one-stop console for the cloud administrator to configure and manage the cloud. The Management Platform also exports various interfaces for the administrator, project manager, developer, and other users, with customizable levels of access and privileges.

To use Hadoop in your private cloud managed by Eucalyptus, you need an image that contains pre-installed Hadoop. If you need, there are some public images containing everything that you need to perform their tasks in your private cloud. This images has a virtual machine format AMI (Amazon Machine Image), due interoperability between Eucalyptus and Amazon. There is links in the Eucalyptus to

pre-packaged virtual machines that are ready to run in your Eucalyptus cloud.

Eucalyptus is an open source implementation for the emerging standard of the EC2 API, designed to simplify the process of building and managing an internal cloud for businesses of any size, thereby enabling companies to create their own self-service infrastructure.

Ubuntu distributions now include the Eucalyptus software core as the key component of the Ubuntu Enterprise Cloud, bringing Amazon EC2-like infrastructure capabilities inside the firewall.

4. Security Issues

Presented as one of the key aspects for the adoption of cloud computing, security has been widely discussed among industries and governments in search of standards for the use of cloud computing that provides information security, including privacy protection, safeguarding the interests security, whether national, in the case of governments, or internal, in the case of industries. (Kundra, 2010).

The decision to adopt or not cloud computing is based on risk, not technology, since the cloud assumes multiple levels of security in information systems, depending on the cloud-sourcing (the supply of resources needed by the business process) model used. Figure 3 illustrates these models, being possible to replace the idea of using government for corporate use.

The proposal, as well as any type of adoption of the cloud, involves security-related issues, which vary according to the deployment model to be chosen, since each has a specific security levels. Address issues relating to deployment models most widely used public and private cloud.

4.1. Issues on Deployment Models: Public and Private

The public clouds bump into problems of security and legal issues. Since you decide for adoption of cloud, it is necessary a starting point for evaluating any cloud service, Damoulakis (2010) suggests the following key issues: Where are my data? How my data are being protected? Who can access or view my data?

Besides this obvious question, Iball (2010) suggests a few more questions: How is the facility secured both physically and digitally? How often they conduct routine checks and audit access to restricted areas by system administrators? What industry standards do your services met?

Iball (2010) also adds that how much more questions are made, the smaller the chances of your data being exposed to unnecessary risks, and more, having the right Service Level Agreement (SLA) is also the key for a successful and long-term relationship with a service provider. SLAs makes the trust between the parties (customers and providers) fortified, what is important because from the SLAs will be defined the crucial details for the implementation of the proposal, personalizing all requirements for security.

Beyond the issues raised by the authors, the Cloud Security Alliance (CSA) launched a guide to assist in the adoption of cloud computing, leading to better understand the issues that must be asked, the currently recommended practices and pitfalls to be avoided. (CSA, 2009).

Also contributing, the Jericho Forum has released a document that which aims to enable stakeholders and business decision-makers, to appreciate the key considerations that need to be taken into account when deciding which parts of their business could be operated in which of the available cloud formations. (Jericho, 2009).

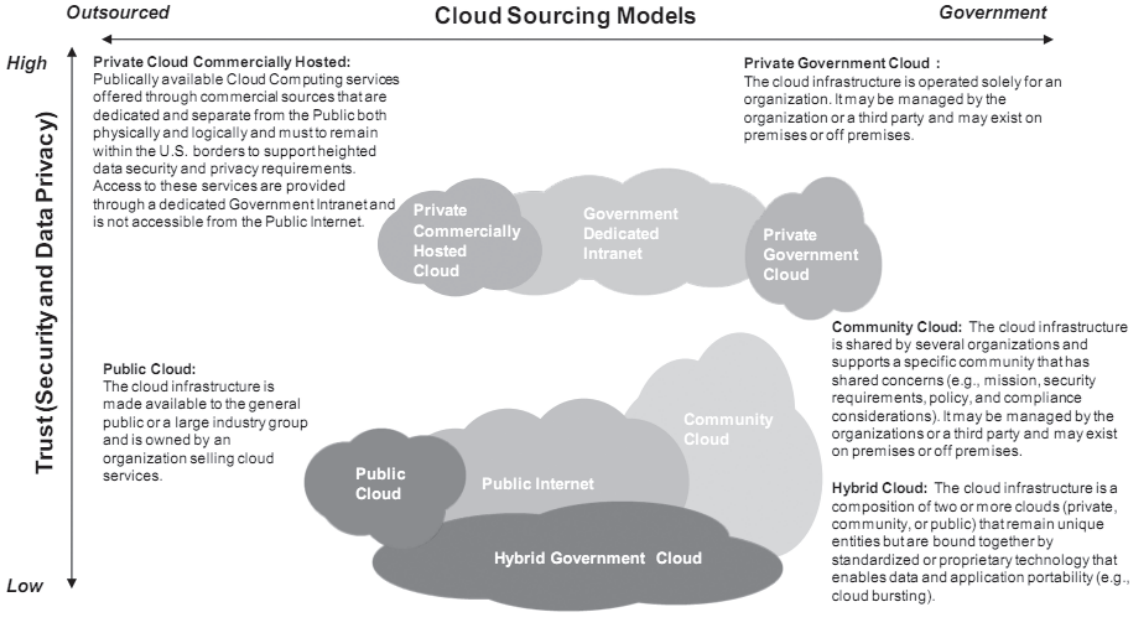

Fig. 3. Cloud Sourcing Models (Kundra, 2010)

Before you set your provider and where your instance runs, you need to focus on the jurisdiction and compliance issues. For example, in the EU the security is treated by law, e.g. European data protection law estipulate that the company, as data controller, is entirely responsible for the security of its own data. The UK's Data Protection Act of 1998 requires that data controllers incorporate security requirement into their contracts with data processors. This means that the onus is on the Small to Medium Size Enterprises (SMEs) to ensure that the cloud computing provider offers the appropriate levels of data security. With UK only data centers, business can be confident of stringent data security practices, covered by UK laws.

The private clouds have a higher benefit compared to the adoption of public clouds. Some issues related to jurisdiction and security mechanisms are already widely understood, since clouds using the private data centers already established in the organizations, running behind your own firewall, which can be managed by the organization itself (which is more secure) or outsourced.

Private clouds enable you to have more control over the issues raised previously in the case of public clouds. The private clouds provide the exact location of where the data is being stored, who are responsible about them, and who has access and all issues of digital or physical protection.

Because of this, the option of implementing the proposal in private cloud becomes more profitable. Anyway, companies still show fear and insecurity, being desirous of moving sensitive data to the cloud. However, the use of cloud computing today is a matter of trust, and studies show that such a system will reach a mature level, ensuring data security.

Confidentiality, privacy, integrity and auditability are still unaddressed issues, and to ensure that the cloud can be used to deliver Forensic as a Service, one should take into account these issues in cases of digital forensic.

According to the report released by Accenture (2010), which makes possible the construction of private clouds is the large number of servers that organizations have, encouraging the establishment for their own use. They also said that 60% of the respondents said they are already using private clouds, often in conjunction with public clouds, i.e. hybrid clouds. According to the report by the end of 2012 the use of private cloud will permeate 77% of companies. For further reading, the full document is in the additional readings.

The private clouds are operated by only one organization, but can be managed by the organization itself or outsourced, can exist as on-premise and off-premise. Our model adopts the form on-premise, which means that the software is installed and run on computers on the premises (in the building) of the organization or person using the software. In a private cloud the owner-user has, in principle, full control over the feature set of the cloud implementation, however there are costs which, cannot be shared with other customers, associated with this increase in control.

For both users of public and private cloud the issue of interoperability between service providers are still causing concern. If there is a change between public cloud providers, private to public or any other settings changes, there are no assurances that the technologies are compatible and will enjoy the features of the principle of interoperability. To address these issues, was created a forum entitled Cloud Computing Interoperability Forum – CCIF (www.cloudforum.org/). A key focus will be placed on the creation of a common agreed upon framework/ontology that enables the ability of two or more cloud platforms to exchange information in a unified way.

The software platform presented as solution of deployment of private cloud from our proposal, Eucalyptus, follows the principles of interoperability with our proposal to implement public cloud using Amazon's services. Eucalyptus supports the popular AWS cloud interface allowing these on-premise clouds to interact with public clouds using a common programming interface. (Eucalyptus System, 2009).

4.2. Software

Traditional models of software systems usually restrict its use to computers where they are running. Because of this, many cloud providers use open-source tools, since commercial tools do not have a good a license model for cloud computing.

Moreover, current forensic tools could not take advantage of cloud environments, with respect to the use of MapReduce computation model, for both is worth discussing the need to re-write code for effective implementation of the model.

Parallel programming is more efficient and can also be used to solve problems in large data sets using non-local resources. When you have a set of networked computers, in our case the cloud, we have a large computational power at hand we can make use of distributed processing and parallel programming using MapReduce to solve the problems involving large data sets.

Many forensic applications can be specified in this model, and examples can be founded in the work of Roussev et al. (2009), citing the use of wordcounter that calculates the number of times a word appears in a text; pi-estimator that calculates an approximation of the value of PI based on Monte Carlo method; grep that is a text search and the bloomfilter application that hashes a file in 4 KB blocks using SHA-1 and inserts them into a Bloom filter.

4.3. Evaluation and validation of results

The purpose of this section is to report the performance achieved by the proposal in its implementation in the public cloud model, adopting the Amazon Elastic MapReduce service in addiction with Amazon Elastic Compute Cloud (EC2) and Simple Storage Service (S3). We utilized an example provided by Hadoop, wordcount program that calculates the

number of occurrences of each word in a text file, featuring a text processing program.

Instances configuration were chosen for implementation are described by Amazon as High-Memory Instances, with the following configurations: 17.1 GB of memory RAM; 6.5 ECU (2 virtual cores with 3.25 ECU each); 420 GB of local instance storage; 64-bit platform, being utilized 2 instances o this type. We have to highlight that according to the Amazon, one EC2 Compute Unit (ECU) provides the equivalent CPU capacity of a 1.0-1.2 GHz 2007 Opteron or 2007 Xeon processor.

We tested wordcount program against 5 files of different sizes (10, 100, 1024, 2048 e 3072 MB each), that contains random text. Results can be appreciated as follow:

- 10 MB – 38 s.;
- 100 MB – 1 min. 26 sec.;
- 1024 MB – 09 min. 58 sec.;
- 2048 MB – 19 min. 51 sec.;
- 3072 MB – 21 min. 12 sec.;

In a single workstation testes (with Core 2 Duo 2 GHz and 2 GB of RAM), our proposal presented better results

in the files with 2 GB – or more – presenting a time diminution of 15 min. 10 sec. which means a time saved around 55%, besides the reduction in the amount of data that approximately 99%.

For better validate our proposal, works such as Peisert, Bishop and Marzullo (2008), aimed at discussing the various forensics systems and situations in which these systems can produce valid and accurate conclusions and in which situation results are suspect. (Peisert et al., 2008)

Another aspect of assessment tools is given by Garfinkel (2010), using standardized data sets, where he argues that the use of these data sets increase scientific assessment of forensic methods, besides the obvious benefits of providing a data set ready and enable a direct comparison between different approaches and tools. There is also an effort by the NIST in creating these data sets, providing to the researchers documented data sets of simulated digital evidence for examination, providing a site that is a repository of images, some produced by the NIST and others are contributions from various organizations (http://www.cfreds.nist.gov/).

Both methods of validating the proposed tools are crucial steps in the further evaluation of this proposed model, thus ensuring integrity and reproducibility of results, in addition to accepting the results as evidence to the jury.

5. CONCLUSIONS

This proposal aims to address one of the challenges of computer forensic, ranked by Garfinkel (2010), where growth in storage capacity of the devices unfeasible the time needed for creation of forensic images, or processing data found.

We presented the use of MapReduce computing model, distributed through the cloud in their public or private form, thus delivering Forensic as a Service - FaaS as a way of solving the challenge of increasing capacity of storage devices and large amounts of data to be analyzed.

The use of cloud computing has been consolidating and giving indications that will be more present in daily life of people and now, we just need that patterns of use, degrees of reliability, privacy and other security and legal issues listed in the text are established.

Legal issues still walk slowly, since there is crimes not covered by current Law yet, but we must move forward while ensuring that such role models may be used in courts in cases involving the elucidation of cybercrimes. This is just one solution to the problem of the amount of data, other solutions may use a different approach, bringing new results with new technologies.

References

[1] Accenture. Cloudrise: Rewards and Risks at the Dawn of Cloud Computing. High Performance Institute. 2011.

[2] Amazon Web Services – AWS. Amazon Elastic MapReduce. Retrieved January 6, 2011, from http://aws.amazon.com/elasticmapreduce/. 2011.

[3] Apache Hadoop. HDFS Architecture Guide. Retrieved February, 11, 2011, from http://hadoop.apache.org/. 2011

[4] Buyya. Rajkumar, Yeo. Chee Shin, Venugopal. Srikumar (2008). Market-Oriented Cloud Computing: Vision, Hype, and Reality for Delivering IT Services as Computing Utilities. 10th IEEE International Conference on High Performance Computing and Communications - 2009, p. 5-13.

[5] Cloud Security Alliance – CSA. Security Guidance for Critical Areas of Focus in Cloud Computing V2.1. 2009

[6] Jim Damoulakis. In Clouds we Trust. IT NOW, 52(2), 11-12. doi:10.1093/itnow/bwq142. 2010.

[7] Jeffrey Dean and Sanjay Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. Sixth Symposium on Operating System Design and Implementation – OSDI 2004. Google Inc.

[8] Eucalyptus System. Eucalyptus Open-Source Cloud Computing Infrastructure - An Overview. Retrieved February 5, 2011, from http://www.eucalyptus.com/pdf/whitepapers/Eucalyptus_Overview.pdf. 2009.

[9] Simson L. Garfinkel. Digital Forensics Research: The next 10 years. Digital Forensics Research Workshop – DFRWS 2010.

[10] John Iball. Don't Cloud Data Security. ITNOW, 52 (2), 14-15. doi: 10.1093/itnow/bwq143. 2010.

[11] Jericho Forum. Cloud Cube Model: Selecting Cloud Formations for Secure Collaboration. Retrieved February 6, 2011 from http://www.opengroup.org/jericho/cloud_cube_model_v1.0.pdf. 2009.

[12] Vivek Kundra. State of Public Sector Cloud Computing. Washington, DC, 2010.

[13] Henry C. Lee, Timothy Palmbach and Marilyn T. Miller. Henry Lee's Crime Scene Handbook. London: Academic Press, 2001.

[14] Carlos E. Marins. Desafios da informática forense no cenário de Cloud Computing. Proceedings of the Fourth International Conference of Computer Science, 2009.

[15] Petter Mell and Tim Grance. The NIST Definition of Cloud Computing. National Institute of Standards and Technology (NIST), Special Publication 800-145, Information Technology Laboratory, 2011.

[16] Kara Nance, Brian Hay, and Matt Bishop. Digital forensics: defining a research agenda. 42nd Hawaii International Conference on System Sciences – 2009.

[17] Daniel Nurmi, Rich Wolski, Chris Grzegorczyk, Graziano Obertelli, Sunil Soman, Lamia Youseff and Dimitrii Zagorodnov. The Eucalyptus Open-source Cloud-computing System. Cloud Computing and Its Application, 2008.

[18] Sean Peisert, Matt Bishop and Keith Marzullo. Computer Forensics in Forensis. 3rd International Workshop on Systematic Approaches to Digital Forensic Engineering – SADFE 2008, 102 – 122. doi:10.1109/sadfe.2008.18.

[19] Regional Computer Forensics Laboratory – RCFL. Annual Report for Fiscal Year 2010. U.S. Department of Justice. Federal Bureau of Investigation. Quantico, VA, 2010.

[20] Vassil Roussev and Golden G. Richard III. Breaking the Performance Wall: The Case for Distributed Digital Forensics. Digital Forensics Research Workshop - DFRWS 2004.

[21] Vassil Roussev, Liqiang Wang, Golden G. Richard III and Lodovico Marziale. MMR: A Platform for Large-Scale Forensic Computing. Fifth Annual IFIP WG 11.9 International Conference on Digital Forensics – 2009.

[22] Kulesh Shanmugasundaram, Nasir Memon, Anubhav Savant and Herve Bronnimann. ForNet: A Distributed Forensics Network. 2nd International Workshop on Mathematical Methods, Models and Architectures for Computer Networks Security, 2009.

[23] Yongping Tang and Thomas E. Daniels. A Simple Framework for Distributed Forensics. 2nd International Workshop on Security in Distributed Computing Systems (SDCS – ICDCSW 2005), 163-169.

[24] Luis M. Vaquero, Luis Rodero-Merino, Juan Caceres and Maik Lindner. A Break in the Clouds: Towards a Cloud Definition. ACM – SIGCOMM 2009, Computer Communication Review, 39(1), 78-85.