Digital Investigation 18 (2016) 96–109

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

A RAM triage methodology for Hadoop HDFS forensics

Petra Leimich [a, *], Josh Harrison [b], William J. Buchanan [a]

[a] Centre for Distributed Computing, Networks, and Security, Edinburgh Napier University, Edinburgh EH10 5DT, UK
[b] Corax Cyber Security, 535 Mission Street, San Francisco, CA 94105, USA

ARTICLE INFO

Article history:
Received 31 October 2015
Received in revised form 11 March 2016
Accepted 16 July 2016
Available online 18 July 2016

Keywords:
Digital forensics
Distributed filesystem forensics
Cloud storage forensics
Hadoop forensics
Triage
RAM forensics
Big data

ABSTRACT

This paper discusses the challenges of performing a forensic investigation against a multi-node Hadoop cluster and proposes a methodology for examiners to use in such situations. The procedure's aim of minimising disruption to the data centre during the acquisition process is achieved through the use of RAM forensics. This affords initial cluster reconnaissance which in turn facilitates targeted data acquisition on the identified DataNodes. To evaluate the methodology's feasibility, a small Hadoop Distributed File System (HDFS) was configured and forensic artefacts simulated upon it by deleting data originally stored in the cluster. RAM acquisition and analysis was then performed on the NameNode in order to test the validity of the suggested methodology. The results are cautiously positive in establishing that RAM analysis of the NameNode can be used to pinpoint the data blocks affected by the attack, allowing a targeted approach to the acquisition of data from the DataNodes, provided that the physical locations can be determined. A full forensic analysis of the DataNodes was beyond the scope of this project.

Introduction

We are at an 'evolutionary point in a new era of the computing environment' (Daryabar et al., 2013). To satisfy the ever-increasing throughput requirements of big data, the use of distributed computing architectures is growing exponentially, and corporate giants such as Facebook, Amazon, and Yahoo! all now use data centres with thousands of nodes holding many petabytes of data. Such data stores pose an attractive target to criminals, and The Cloud Security Alliance (2014) name 'Data breach' and 'Data loss' as two primary threats to cloud computing, while the ISC^2 White Paper IX (2013) identified seven main concerns in relation to cloud security that include data loss, forensic readiness, and uninterrupted availability. It is clear from these concerns that despite the widespread adoption of distributed and cloud computing, there is uncertainty as to whether the technology can handle a data breach scenario. Garfinkel (2010) had already foreseen the coming challenges, with the bleak prognosis that digital forensics was at the end of its golden age, entering a time of crisis due to expanding technologies and their technical, business-centric and legal challenges.

Apache Hadoop is among the most implemented distributed computer architectures for storing and processing big data. Used by Internet giants and SMEs alike (the latter often through outsourced services), Hadoop has been transformative in the business sphere with an estimated 76% of Fortune companies implementing the technology by 2015 (Business insider, 2014). Thus we have selected a Hadoop implementation to propose and test a forensic methodology that exemplifies how the above challenges can be addressed through the use of live RAM forensics to facilitate targeted data acquisition.

We regard the contribution of this paper to be the following:

i. Offer insight into aspects of Hadoop HDFS architecture and how they affect forensic analysis;

* Corresponding author.
E-mail address: p.leimich@napier.ac.uk (P. Leimich).

http://dx.doi.org/10.1016/j.diin.2016.07.003
1742-2876/

ii. Propose a tailored variation of cloud forensic methodology, based on earlier work by various authors, that is based on the findings of contribution i and thus applicable to Hadoop HDFS data breach scenarios (see Section Design and Fig. 1);
iii. Provide a case study that evaluates the feasibility of our methodology (contribution ii) with particular focus on the steps that afford initial triage for data acquisition.

Background

Complex business and legal demands create impediments to 'in-cloud' forensics that add to the problem of traditional forensic approaches being rendered infeasible by the sheer volume of data, however the question of forensic readiness has not yet been answered. The need for a forensic methodology scalable to the big data age is apparent (Quick and Choo, 2014).

In 'traditional' digital forensic investigations well-established guidelines and methodologies, such as the ACPO guidelines (Association of Chief Police Officers) and DFRWS guidelines (Digital Forensic Research Workshop), are used to safeguard the validity and integrity of evidence and the investigative process as a whole. Traditional methodologies utilise 'dead' acquisition techniques as a means of evidence gathering in which identical bit-to-bit images are produced. However, given that indexing speed decreases as the volume of data increases (Lee and Hong, 2011) this approach is not well suited for big data forensic scenarios such as in a Hadoop cluster. Writing to four external devices simultaneously with a transfer rate of over 6 GB/min, it would take 28 days to produce a bit-to-bit image of one petabyte of data (Fowler, 2012). As Fowler (2012) underlines, ideally this image should not be examined directly, but instead used as a master image from which a further copy should be produced for examination, to avoid the risk of irrevocably contaminating the image during an investigation. When factoring in the imaging of the image, the acquisition stage alone of a petabyte of data is 56 days (Fowler, 2012).

When considering forensics, it is important to contextualise Hadoop's adoption within the business sphere, as further challenges become apparent. Although Hadoop can be run in-house at a company's own data centre, SMEs will typically not have the facilities or the administrative capacities to maintain their own cluster. In these instances, Hadoop will be used as a Platform as a Service (PaaS) through cloud service providers (CSPs) such as Amazon's EC2 (Grispos et al., 2012). If a data breach were to occur in these instances then the acquisition stage of the forensic investigation also takes on further legal and ethical issues; namely multi-tenancy – multiple clients sharing access to a CSP's data nodes (Barrett and Kipper, 2010; Martini and Choo, 2012), and organisational – the involvement of a third party (the CSP) in the investigation (Ruan et al., 2011). A further legal consideration arises when considering that a CSP's cluster may physically reside in a different country from the breached client, meaning that both parties are governed by different legislation and jurisdictions (Spyridopoulos and Katos, 2011).

Blanket dead acquisition is infeasible when these considerations are made. Indeed, even if an organisation maintains its own datacentre, rendering the legal considerations less applicable, the lengthy process of imaging all of the nodes in the cluster would still cause undesirable downtime resulting in a loss of business (Cho et al., 2012). It is therefore apparent that methodologies such as the ACPO and DFRWS guidelines are unworkable for big data storage environments (Grispos et al., 2012; Hegarty et al., 2012; Lallie and Pimlott, 2012; Martini and Choo, 2012; Poisel et al., 2013), and a new set of both technical and procedural guidelines need to be established to deal with the complex challenges highlighted (Cho et al., 2012; Martini and Choo, 2012).

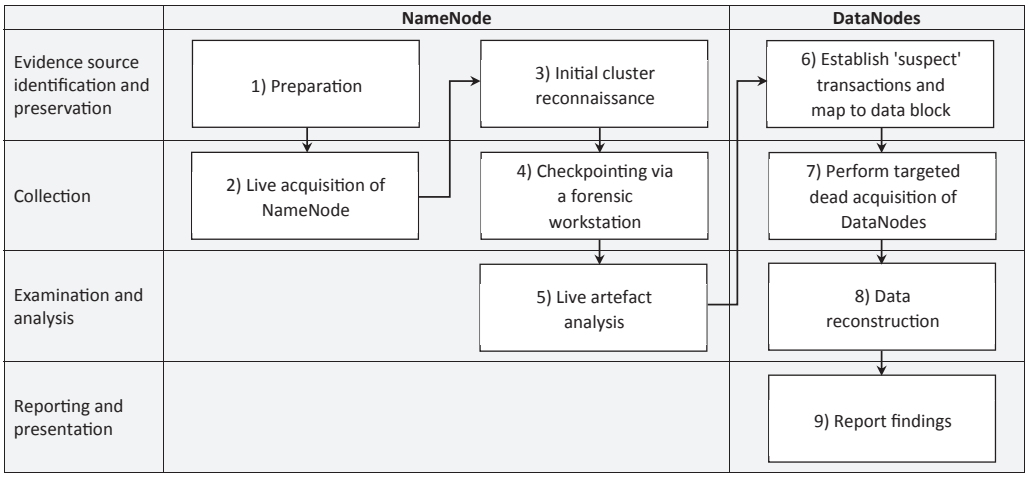

Fig. 1. Proposed forensic methodology. The nine phases are described in Section Design and applied to a test scenario in Section Forensic process – application of methodology. Labels on the left match those used by Martini and Choo (2012, 2014a) for comparison, while labels at the top identify the corresponding nodes.