Contents lists available at ScienceDirect

Digital Investigation

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

Digital forensics as a service: Game on

H.M.A. van Beek*, E.J. van Eijk, R.B. van Baar, M. Ugen, J.N.C. Bodde, A.J. Siemelink

Netherlands Forensic Institute, Laan van Ypenburg 6, 2497 GB The Hague, The Netherlands

ARTICLE INFO

Article history:
Received 12 March 2015
Received in revised form 8 July 2015
Accepted 11 July 2015
Available online xxxx

Keywords:
Distributed systems
Digital forensics
Big data
XIRAF
HANSKEN

ABSTRACT

The big data era has a high impact on forensic data analysis. Work is done in speeding up the processing of large amounts of data and enriching this processing with new techniques. Doing forensics calls for specific design considerations, since the processed data is incredibly sensitive. In this paper we explore the impact of forensic drivers and major design principles like security, privacy and transparency on the design and implementation of a centralized digital forensics service.

Introduction

Many papers in the field of digital forensics start with the observation that the size of digital material increases, that the complexity and diversity of the digital evidence grows and that more advanced techniques are needed to be able to keep up with the evolving digital society.[1]

Since December 2010, the Netherlands Forensic Insitute has been using a service-based approach for processing and investigating high volumes of seized digital material: Digital Forensics as a Service (DFaaS) (van Baar et al., 2014). This service is called XIRAF (Bhoedjang et al., 2012).

Now, four years later, this approach has become a standard for hundreds of criminal cases and over a thousand investigators, both in The Netherlands and abroad. After having processed over a petabyte of data, we have experienced the impact of the XIRAF system and the paradigm shift it is causing (van Baar et al., 2014). XIRAF started in 2006 as a scientific research project aimed at identifying and developing techniques for automating (parts of) the data analysis process. XIRAF was never meant to be an operational system for processing petabytes of data and providing access to over a thousand investigators. As a result, design decisions taken during the development of XIRAF leave room for improvement.

In the beginning of 2012, we started working on the successor of XIRAF, named HANSKEN. This work consisted of defining design principles, building a proof of concept (PoC) based on the new principles and ideas, making design decisions based on the principles and PoC and building a production version to replace XIRAF. This paper provides an overview of the major design decisions that form the foundation for the HANSKEN solution for providing digital forensics as a service.

A lot of challenges arise when building a system to provide insight in petabytes of different types of data. Especially when integrity and confidentiality of the data are crucial. While it is tempting to focus on developing new techniques and building bigger and faster systems, boundaries need to be established in which such a system can operate. Without these boundaries, major risks of data breaches and leaks of sensitive information exist.

* Corresponding author.
E-mail address: harm.van.beek@nfi.minvenj.nl (H.M.A. van Beek).

[1] http://apmdigest.com/gartner-top-10-strategic-technology-trends-for-2013-big-data-cloud-analytics-and-mobile, visited March 11, 2015.

http://dx.doi.org/10.1016/j.diin.2015.07.004
1742-2876/© 2015 Elsevier Ltd. All rights reserved.

In Section 2, the reasons why a forensic big data solution is desirable are described (forensic drivers), as well as the motivation for the boundaries of such a system (design principles). Section 3 contains considerations for how the forensic drivers and design principles affect the chosen solutions. Section 4 describes different solutions we have implemented in Hansken in order to cope with the considerations while still being able to do digital forensics. A lot of work has been done in the field of forensics and big data, even though the term was not yet used in most related work. Section 5 discusses different topics related to big data and how we see them match with a forensic big data platform like Hansken. Finally, we draw conclusions in Section 6.

Motivation

Business needs provide the main reasons for developing and providing a centralized system for doing large scale forensic data analysis. First of all, cost reduction asks for automating parts of the extraction and analysis process. Here, the economies of scale apply (Armbrust et al., 2010). Secondly, centralization makes it possible to standardize forensic data extraction and analysis and increase its quality. All this is explained in detail below.

As mentioned, the Netherlands Forensic Insitute has been providing Digital Forensics as a Service to the Dutch law enforcement organizations since December 2010. Fig. 1 shows the procedure of handling digital forensic cases using this approach.

On the right, there are detectives and analysts that have questions related to information presumably available in the digital material shown on the left. To guarantee forensic integrity, forensic images are needed (van Baar et al., 2014; Kohn et al., 2013), so the first task is to create these forensic copies of the digital devices. The images are copied to a central storage and processed using a standard set of tools. We call this the *extraction process*. The applied tools range from tools that analyze file systems, extract files, carve unallocated space and create full text indexes, to tools that parse chat logs, browser history and e-mail databases. The results of these tools, i.e. the extracted *metadata*, are stored in a centralized database. The combined data and metadata of this process are referred to as *traces*, e.g. an e-mail, chat message or zip archive. After storing these traces, they can be queried using multiple methods: detectives can log on using a web browser and query the traces by applying filters and text searches. Digital investigators can use the programming interface to run automated tools and scripts written in their favorite programming language. Analysts may want to retrieve all information and analyze the results using data visualization tools, integrate additional data sources or build a network of contacts, for example. This makes it possible to identify, classify, organize and compare the traces within seconds, based on hypotheses and questions the investigators have. This can be done at any time during the investigation.

To support this process, the next paragraphs discuss the three drivers, eight design principles and our two ways of looking at data in the system. This defines the scope within which we designed Hansken.

Forensic drivers

Our main goal is to provide a service that processes high volumes of digital material in a forensic context and gives easy and secure access to the processed results. We identify three main forensic drivers: minimization of the case lead time, maximization of the trace coverage and specialization of people involved. These forensic drivers are the reasons for building a big data forensic platform.

Minimize case lead time

Generally, the first 48 hours of an investigation are critical to an investigation (U.S. Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention, May 1998; Joyce, 2012; Wikipedia). Traditionally, results from digital investigations are not available in these first days. Traces found in digital material are therefore often used for validating hypotheses instead of forming them. In an ever increasing digital society, digital evidence becomes more and more key evidence. This makes it unacceptable to exclude digital material from the initial response: digital material must be available to the investigation team in those 48 hours. In this context, available does not only mean that investigators have access to the data, but also that they have tooling at their disposal for finding relevant traces. Examples are high performance filtering tools based on trace details or keywords and visualization tools for presenting search results.

To give access to the traces within 48 hours, processing of the seized material must be automated. This has high impact on the way digital material should be handled (van Baar et al., 2014). Furthermore, the results of this automated process must be made available to the investigation team directly and not to specialized digital investigators. This is discussed below. Since investigation teams can be scattered over multiple locations, access to the data and extracted traces should not be limited by e.g. building or department boundaries.

To speed up the investigation, detectives should be able to annotate or tag interesting traces or traces they do not understand. Other detectives and digital investigators must have access to the annotation so that they can act on it.

Maximize coverage

Seized material varies wildly, both in types of devices (hard drive, volatile memory or mobile phones), but also in file systems and file formats contained in forensic copies made from these devices. This is caused by simple things like software upgrades and the availability of new devices and new software for existing devices. This variation requires constant attention to make sure that the traces contained in the data are extracted. To keep up with software upgrades as well as counteract the ever increasing sophistication in technology used by suspects, increasing sophistication in the trace extraction tools is needed.

Processing petabytes of data in a central environment means processing a large variety of images. This requires the tools to process many different file formats, database formats and applications. Centralized processing of more and more data results in increased insight in the coverage