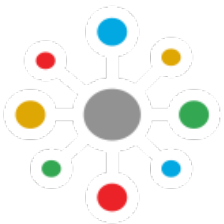
Big Data and Cloud Computing in Digital Forensics: Trends and Challenges

Marek Rychlý

Department of Information Systems
Faculty of Information Technology
Brno University of Technology
(Czech Republic)



TARZAN Project Meeting
10th September, 2019

Outline

Digital Forensics Workflow: Overview

Identification	Preservation	Collection	Examination	Analysis	Presentation	Decision
Event/Crime Detection	Case Management	Preservation	Preservation	Preservation	Documentation	
Resolve Signture	Imaging Technologies	Approved Methods	Traceability	Traceability	Expert Testimony	
Profile Detection	Chain of Custody	Approved Software	Validation Techniques	Statistical	Clarification	
Anomalous Detection	Time Synch.	Approved Hardware	Filtering Techniques	Protocols	Mission Impact Statement	
Complaints		Legal Authority	Pattern Matching	Data Mining	Recommended Countermeasure	
System Monitoring		Lossless Compression	Hidden Data Discovery	Timeline	Statistical Interpretation	
Audit Analysis		Sampling	Hidden Data Extraction	Link		
Etc.		Data Reduction		Spacial		
		Recovery Techniques				

(adopted from “A Road Map for Digital Forensic Research, DFRWS USA, 2001”)

Digital Forensics Workflow: Phases & Activities

1. Identify
2. Acquire: Preservation & Collection
   (i.e., to get and to store evidence)
3. Analyse: Examination & Analysis
   (i.e., to extract and to process information)
4. Report: Presentation & Decision
   (i.e., to measure/visualize and to prove/reject)

- The steps should be done with a minimal latency.
- The latency is affected by
  - the volume of the forensic data,
  - the performance of the forensic activities.
- Big Data & Cloud Computing may increase both the volume and the performance.

The “Acquire” Phase: Volume ⇒ Big Data

- Investigators need to depend on organizations to collect data.
  (providers of cloud services, data centers, distributed networks, etc.)
- For example, a forensic evidence in a data center:
  thousands of decentralized machines hosting PB/ZB of data.
  (physical inaccessibility of storage, various jurisdictions, confidential data)
- Difficult to collect the evidence from such infrastructure.
  (the current digital forensics tools mostly designed for personal owned computing system, where the volume of hard drive can be a few terabytes at most)

Possible approaches:

- To narrow focus on usage residues of dist. services, not on the services.
  (e.g., logs, cached items, data fragments, etc.)
- To employ data reduction techniques early in the acquisition phase.
  (distributed agents to gather data of specific types, origin, etc.)

The “Acquire” Phase: Variety, Velocity and Veracity

- Wide variety of structures of big data.
  (e.g., no common standards for the data format for IoT devices)
- Intensive data-flows prevent effective storing & batch-processing.
  (it is necessary to narrow a context or a timescale of the captured data)
- Low quality and accuracy of big data.
  (missing data/offline nodes, desynchronized clocks among nodes, etc.)

Possible approaches:

- To distribute the forensics procedures to data/computation providers.
  (one distributed, not several isolated, process with cooperation of the providers)
- Dynamic on-demand collection of data based on needs of later phases.
  (acquire only those data that can be processed in the late phases)

A Sample Technical Concept for the “Acquire” Phase

- Multi-level object storage of unstructured forensic data.
  (lower levels: full forensic images; higher levels: reduced data-sets; to store unstructured data objects w/GUIDs & metadata attrs., e.g., OpenStack Swift)
- Hadoop File-system (HDFS) and NoSQL databases for working data/intermediate results.
  (distributed, fast, and inexpensive, but not much reliable storage)
- Distributed agents (Akka) & tasks (MapReduce) to process data.
  - on-demand acquisition from external data sources,
    (“edge computing” at third parties; access to “approved” forensic images[1])
  - autonomous extraction and tagging of the data by their metadata.
    (to pre-process the data and build inputs for the later phases)

[1] e.g., an integration of AFF4 repository, EWF/E01 and DD images, etc.

The “Analyse” Phase: Distribution

- A need for an extremely costly local infrastructure.
  (to provide the desired functionality in a reasonable time period)
- (Correlation) Analyses on wide data sets require special tools.
  (it is not possible to do such analyses locally)
- Live-feed analyses cannot be done in isolated environments.
  (the live-analyses are required to process high-velocity Big data)

Possible approaches:

- Providing “Forensics-as-a-Service” as a cloud service.
  (a cost effective and easily scalable cloud-based architecture)
- Utilization of framework for distributed machine learning algorithms.
  (Spark MLib, Mahout, etc.)

The “Analyse” Phase: Interpretation

- Simple verification and ad-hoc queries may not be sufficient.
  (not efficient enough to see the context, connections and correlations in data)
- Missing or isolated (by different tools) special view on the data.
  (e.g., resource & time-line view, network analysis, content exploration, etc.)
- Difficult to customize analytical processes to particular domains.
  (to create domain models tailored to needs of individual investigation cases)

Possible approaches:

- Building semantics models of forensic data (what, when, where, who).
  (Semantic Web – ontologies & RDF graphs of facts, networks, time-lines, etc.)
- Semantic and temporal queries of forensic data/information/knowledge.
  (e.g., “give me all web-visits that were followed in two minutes by accessing cloud storage services or creation of BitTorrent files and eventually by creation of at least 100 JPEG files during one minute”)

A Sample Technical Concept for the “Analyse” Phase

- Cloud-based service for forensic data repository and analysis.
  (private edge cloud, i.e., non-public infrastructure w/integrated data providers)
- Automatic analysis of reduce data (both E1 to L1 and from L1[2]).
  (complex event processing by Flink engine & Plaso generator)
- Mapping of the events into RDF/semantic web.
  (GrpahX, Halyard; including customized ontologies)
- Semantic data querying & reasoning to derive new facts.
  (SPARQL, Jena Elephas)

[2] E1/L1 is the first generation of evidence/reduced data, respectively

Summary

- Big Forensic Data should be processed on Big Data platforms.
  (Hadoop stack, Spark, Flink, etc.)
- Advanced end users can be able to use those platforms directly.
  (SQL-like query languages, build apps from predefined components, etc.)
- Moving into cloud-based architectures – “Forensics-as-a-Service”.
  (probably not yet publicly available; why?)