Big Data Threat Landscape and Good Practice Guide

JANUARY 2016

www.enisa.europa.eu European Union Agency For Network And Information Security

About ENISA

The European Union Agency for Network and Information Security (ENISA) is a centre of network and information security expertise for the EU, its member states, the private sector and Europe's citizens. ENISA works with these groups to develop advice and recommendations on good practice in information security. It assists EU member states in implementing relevant EU legislation and works to improve the resilience of Europe's critical information infrastructure and networks. ENISA seeks to enhance existing expertise in EU member states by supporting the development of cross-border communities committed to improving network and information security throughout the EU. More information about ENISA and its work can be found at www.enisa.europa.eu.

Authors

This report was authored by a group of experts, and edited by ENISA:

Authors: Ernesto Damiani (CINI), Claudio Agostino Ardagna (CINI), Francesco Zavatarelli (CINI), Evangelos Rekleitis (ENISA), Louis Marinos (ENISA).

Editor: Evangelos Rekleitis (ENISA)

Contact

For contacting the authors please use isdp@enisa.europa.eu

For media enquiries about this paper, please use press@enisa.europa.eu.

Acknowledgements

We would like to express our gratitude to the working group supported ENISA work for their valuable contribution to this report, including: W.M.P van der Aalst, Eindhoven University of Technology, The Netherlands and Panayotis Kikiras, AGT Group, Germany. Acknowledgement should also be given to ENISA colleagues who helped in this project.

Table of Contents

Executive Summary 5

1. Introduction 7

1.1 Policy context 8

1.2 Scope 8

1.3 Target audience 8

1.4 Methodology 9

1.5 Structure of this document 9

2. Big Data Environments 10

2.1 Big Data architecture 10

3. Big Data assets 12

3.1 Big Data asset taxonomy 12

3.2 Big Data asset categories 14

4. Big Data threats 17

4.1 ENISA threat taxonomy 17

4.2 Mapping threats to Big Data assets 19

4.2.1 Threat Group: Unintentional damage / loss of information or IT assets 20
4.2.2 Threat Group: Eavesdropping, Interception and Hijacking 23
4.2.3 Threat Group: Nefarious Activity/Abuse 23
4.2.4 Threat Group: Legal 27
4.2.5 Threat Group: Organisational threats 29

5. Threats agents 30

6. Good practices 33

7. Gap analysis 43

Annex A: Full list of Big Data taxonomy 48

Annex B: Full Big Data asset taxonomy structure 53

Annex C: Full list of threats affecting Big Data 54

Annex D: Full Big Data threat taxonomy structure 56

Annex E: Big Data analytics for security 57

Annex F: Summary of threat taxonomies 61

Executive Summary

The term Big Data is often used loosely to designate the palette of algorithms, technology and systems employed for collecting data of unprecedented volume and variety, and extracting value from them by massively parallel computation of advanced analytics. The sources of Big Data are many and diverse. Distributed multimedia sensors on the Internet of Things, mobile telecommunication devices and networks, distributed business processes, and Web-based applications are all candidate data providers/generators. As Big Data usage has increased over the years, the various algorithms, technologies, and systems are gradually reaching a level of development and maturity suitable for widespread adoption.

Experience has shown that Big Data applications can provide a dramatic increase in the efficiency and effectiveness of decision-making in complex organizations and communities[1,2]. It is expected that it will constitute an important part of a thriving data-driven economy, with applications ranging from science[3] and business to military and intelligence[9]. However, besides its benefits or in some cases because of them, Big Data also bears a number of security risks. Big Data systems are increasingly becoming attack targets by threat agents, and more and more elaborate and specialized attacks will be devised to exploit vulnerabilities and weaknesses.

This Threat Landscape and Good Practice Guide for Big Data provides an overview of the current state of security in the Big Data area. In particular, it identifies Big Data assets, analyses exposure of these assets to threats, lists threat agents, takes into account published vulnerabilities and risks, and points to emerging good practices and new researches in the field. To this aim, ongoing community-driven efforts and publicly available information have been taken into account.

The study analyses threats to all identified Big Data asset classes. Highlights include:

- Big Data threats include, but are not limited to, threats to ordinary data. The high level of replication in Big Data storage and the frequency of outsourcing Big Data computations introduce new types of breach, leakage and degradation threats that are Big Data-specific.
- Big Data is having significant privacy and data protection impacts. The creation of links at data collection (a.k.a. “ingestion”) time is a key requirement for parallelization – and therefore performance - of Big Data analytics, but the additional information it creates may increase the impact of data leakages and breaches.
- The interests of different asset owners (e.g., data owners, data transformers, computation and storage service providers) in the Big Data area are not necessarily aligned and may even be in conflict. This creates a complex ecosystem where security countermeasures must be carefully planned and executed.
- As in many other areas of ICT, starting to apply basic privacy and security best practices would significantly decrease overall privacy and security risks in the Big Data area. At this still early stage of this emerging paradigm, embracing the Security-by-default principle can prove to be both highly

[1] http://data-informed.com/use-analytics-to-improve-operations-and-energy-efficiency/, accessed November 2015.
[2] http://www.zdnet.com/article/big-data-is-a-competitive-advantage-companies-can-no-longer-ignore/, accessed November 2015.
[3] http://knowledgent.com/whitepaper/big-data-analytics-life-sciences-healthcare-overview/, accessed November 2015.

practical and beneficial; as opposed to the cost and effort required to provide ad hoc solutions later on.

This guide finally provides a gap analysis presenting a comparison between identified Big Data threats and identified Big Data countermeasures. In this context, the lack of current Big Data countermeasures and pressing needs in the development of next-generation countermeasures are discussed. In particular, the question arises of the trend of current countermeasures of adapting existing solutions against traditional data threats to the Big Data environments, mostly focusing on the volume of the data. This practice mainly targets scalability issues and clearly does not fit the Big Data peculiarities (5V- Volume, Variety, Value…) resulting in partial and ineffective approaches. A set of recommendations for next-generation countermeasures concludes the guide. Among these recommendations, we remark i) to depart from current approaches for traditional data, defining Big Data-specific solutions, ii) to identify gaps and needs for current standards, planning the definition of standardization activities, iii) to focus on training of specialized professionals, iv) to define tools for security and privacy protection of Big Data environments, v) to clearly identify Big Data assets simplifying the selection of solutions mitigating risks and threats.

Aligning to its mandate ENISA published two more reports studying the impact of Big Data in the more specialized areas of data protection and privacy ("Privacy by design in big data"[4]) and critical infrastructures[22].

[4] https://www.enisa.europa.eu/activities/identity-and-trust/library/deliverables/big-data-protection/at_download/fullReport, accessed December 2015.

1. Introduction

In this reports ENISA elaborates on threats related to Big Data, a technology that has gained much traction in recent years and is expected to play a significant role affecting various aspects of our sociatey, ranging from health, food security, climate and resource efficiency to energy, intelligent transport systems and smart cities[5]. The European Commission has acknowledged the potential impact of Big Data in a "*thriving data-driven economy*" by outlining a strategy on Big Data[6]. According to estimates, the value of just personal data of EU citizens has *"the potential to grow to nearly €1 trillion annually by 2020"*[7] (sic). It is thus conceivable that data will continue to be a significant economic drive. But also in science and research Large and nowadays Big Data continue to proliferate and many agencies and institutions in Europe and around the globe have or are planning to launch Big Data projects to facilitate scientific data analysis and exploitation[8]. Big Data technologies are also being used in military applications; such as fighting terrorism; assisting in combat; gathering and analysing intelligence from heterogeneous sources, including battlefield data and open sources[9]. In addition, many existing data intensive environments have in recent years adopted a Big Data approach. To name just a few examples, Facebook[10] is thought to store one of the biggest datasets worldwide, storing more than 300 petabytes of both structured and unstructured data; Twitter recently decided to tap directly into its own raw data using Big Data analytics[11] and the world's telecommunications capacity was already by 2007 near 65 Exabytes (without signs of this trend declining)[12]; straining existing storage and analytic processes and technologies.

Given that Big Data approaches make use of extremely novel and high tech ICT systems, with little time to mature against cyber-attacks it is not suprising that attacks are showing an increased trend in both number, sophistication and impact. But because of the loose use of definitions and the unwillingness of affected organizations to disclose attack data, accurate estimates are not easy to come up with. Additionally, as more and more businesses and organizations venture into the Big Data field, attackers will have more incentives to develop specialized attacks against Big Data. Somewhat paradoxically, Big Data approaches can also be used as a powerful tool to combat cyber threats by offering security professionals valuable insigts in threats and incident management.[13]

---

[5] See https://ec.europa.eu/digital-agenda/en/towards-thriving-data-driven-economy, accessed December 2015.
[6] "Towards a thriving data-driven economy". Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions, http://ec.europa.eu/digital-agenda/en/news/communication-data-driven-economy, accessed December 2015.
[7] European Commission memo, "Progress on EU data protection reform now irreversible following European Parliament vote", Strasbourg, 12 March 2014 http://europa.eu/rapid/press-release_MEMO-14-186_en.htm, accessed December 2015.
[8] http://byte-project.eu/10-big-data-initiatives-an-insight-into-the-big-data-landscape/, accessed December 2015.
[9] Defense One, "Harnessing Big Data to Protect the Nation", http://www.defenseone.com/reports/harnessing-big-data/122177/, accessed Nov 2015.
[10] See https://www.facebook.com/notes/facebook-engineering/presto-interacting-with-petabytes-of-data-at-facebook/10151786197628920, accessed December 2015.
See [11] https://blog.gnip.com/twitter-data-ecosystem/, accessed December 2015.
[12] M. Hilbert and Pr. López , "The World's Technological Capacity to Store, Communicate, and Compute Information". Science (journal), 332(6025), 60-65, (2011)
[13] Recent ENISA Threat Landscapes (2013, 2014, 2015) considered the Big Data field as an emerging threat area both because it is a valuable asset and as such is being targeted by cyber-attacks and because it has the potential to become a very powerful tool for security professionals.

Being an ENISA deliverable in the area of Threat Landscape, this report constitutes a detailed threat assessment in the area of Big Data, based on input from the ENISA Threat Landscape activities. The rationale behind this piece of work is to "deepen" the generic threat assessment by taking into account the specificities of Big Data.

1.1 Policy context

Threat analysis and emerging trends in cyber security are an important topic in the Cyber Security Strategy for the EU[14]. Moreover, the new ENISA regulation[15] highlights the need of analysing current and emerging risks and dictates that "*the Agency, in cooperation with Member States and, as appropriate, with statistical bodies and others, collects relevant information*". More specifically, it is stated that it should "*enable effective responses to current and emerging network and information security risks and threats*".

To this end the ENISA Work Programme 2015[16] included this study on "*Big Data Threat Landscape and Good Practice Guide*" as one of this year's deliverables ("*WPK1.1-D2: Risk Assessment on two emerging technology/application areas*" that focuses on Big Data).

The report aims to identify emerging trends in cyber-threats and to provide a concise state of the art analysis of the cyber threat and security issues of Big Data; consolidating existing and open literature and available information, and contributing to a cyber security public and private initiatives by addressing industry concerns in the area..

1.2 Scope

This report contributes to the definition of a threat landscape, by providing an overview of current and emerging threats applicable to Big Data technologies, and their associated trends. Several Big Data definitions exist in the literature and the area is constantly being shaped by advantages in methods, tools, and new applications, thus it is not possible to take into consideration all Big Data systems. The research done focuses on assets, threats and controls applicable to prominent, important and/or widely used Big Data systems.

The goal is to deepen our understanding of the threats that affect Big Data and to provide good practices and recommendations for those threats that are considered important or emerging.

1.3 Target audience

It is expected that this report will be useful for performing detailed Risk Assessments (RA) and Risk Management (RM) by Big Data providers and operators according to their particular needs and for Big Data consumers in drafting their SLAs. The asset and threat taxonomies presented here are to be expanded by asset owners, based on the particular Big Data system instantiation at hand, before being used as input to RA/RM and cyber threat exposure analysis.

---

[14] See http://ec.europa.eu/digital-agenda/en/news/eu-cybersecurity-plan-protect-open-internet-and-online-freedom-and-opportunity-cyber-security, accessed December 2015.

[15] Regulation (EU) No 526/2013 of the European Parliament and of the Council of 21 May 2013 concerning the European Union Agency for Network and Information Security (ENISA) and repealing Regulation (EC) No 460/2004 Text with EEA relevance, http://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32013R0526, accessed December 2015.

[16] "ENISA Work Programme 2015", https://www.enisa.europa.eu/publications/programmes-reports/, accessed December 2015.

Moreover, the presented Big Data threat landscape will be of use to policy-makers for understanding the current state of threats and respective mitigation practices and measures in the area.

Further, the extensive research of relevant existing literature in Big Data security and threat research means this study will be of particular interest to researchers and institutions working in the field.

1.4 Methodology

This study and its outcome are based on desk research and review of conference papers, articles, technical blogs and a variety of other open sources of information relevant to Big Data. This report identifies the majority of sources consulted; the details of all documentary sources consulted during this study are available on request. More than one hundred documentary sources were identified through a number of search methods, including specialist search engines for academic sources and journal articles. The sources collected are all in English.

The overall work went through a three-step process as follows: The first step "Information collection" was about the identification and collection of relevant information, in particular the assets and threats. The second step "Assessment, Guidelines and Gap Analysis" performed an analysis about the collected information to identify current and emerging trends and then elaborated countermeasures in a Big Data scenario. The third step "Good practices definition" was focused on findings, current practices, and needs that formalized the Big Data threat landscape report.
A final note, all referenced web resources were last accessed in November 2015.

1.5 Structure of this document

The structure of document is as follows: in section 2 we define Big Data and describe an abstract architecture upon which the study is based; in section 3 we present an asset taxonomy for Big Data; in section 4 we identify threats against Big Data, based on the threat taxonomy used by ENISA in "*Threat Landscape and Good Practice Guide*" reports, and map these threats to Big Data assets; in section 5 we consider which threat agents are more relevant to Bog Data attacks; in section 6 we present a set of recommendations and good practices for Big Data; we conclude in section 7 with a gap analysis.

In addition 6 annexes are provides at the end of the report.

Annex A contains the Big Data asset taxonomy in full depth; including all identified asset groups, asset types, assets and asset details.

Annex B contains the detailed Big Data asset taxonomy diagram.

Annex C contains the Big Data threat taxonomy in full detail; including all identified threat groups/types correlated to threat agents and affected Big Data assets.

Annex D contains the detailed Big Data threat taxonomy diagram.

Annex E contains a concise presentation on how Big Data analytics can assist security professionals in analysing threats and attacks and detecting intrusion and fraud cases.

Annex F contains a summary of existing threat taxonomies, which were used along with ENISA's threat taxonomy to drive this study.

2. Big Data Environments

The term Big Data[17] describes the vast amount of data in our information-driven world. In a 2001 research report[18] and related lectures, the META Group (now Gartner) defined the data growth challenges and opportunities as being three-dimensional, i.e. ***increasing Volume*** (amount of data), ***Velocity*** (speed of data in and out), and ***Variety*** (range of data types and sources). Gartner, and then the industry, used this "3Vs" model for describing Big Data: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization.[19]". Additionally, some new Vs have been added by some organizations to further define Big Data: "***Veracity***" (data authenticity since the quality of captured data can vary greatly and an accurate analysis depends on the veracity of source data), "***Variability***" (data meaning is often changing, and the data can show inconsistency at times, and this can hamper the process of handling and managing the data effectively) and "***Value***" (the potential revenue of Big Data). This being a developing field, several other alternative or complenetary definitions have been proposed, in an effort to capture different nuances attributed to Big Data; such as its evolutianary nature: "*datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze*." (sic)[20]. Given that the field is still not mature, for the purposes of this report we take into account the different ways Big Data is defined.

While a great scientific opportunity exists with Big Data, this growth is outpacing the technological advances in computational power, storage, analysis and analytics. Furthermore a real concern is arising about the security of this massive amount of digital information, the data protection and privacy issues[4,21,] and the protection of the (critical) infrastructure supporting it[22].

2.1 Big Data architecture

The architecture in **Error! Reference source not found.** is a high-level conceptual model that facilitates the discussion of security requirements in Big Data and introduces the terminology used in this report. It does not represent the system architecture of a specific Big Data system, nor it is tied to any specific vendor products, services, or reference implementation, but rather it is a tool for describing some common Big Data components; i.e. the Big Data environment. In our vision the notion of Big Data architecture can be

[17] A short history of this term can be found in "A Very Short History of Big Data" (http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/). The first article in the ACM digital library to use the term "Big Data" was "Application-controlled demand paging for out-of-core visualization", written by Michael Cox and David Ellsworth and published in the Proceedings of the IEEE 8th conference on Visualization (1997).

[18] Laney, Douglas. "3D Data Management: Controlling Data Volume, Velocity and Variety" (PDF). Gartner. Issued: 6 February 2001. http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf, accessed December 2015.

[19] See Gartner IT glossary at http://www.gartner.com/it-glossary/big-data, accessed December 2015.

[20] J. Manyika "Big data: The Next Frontier for Innovation, Competition, and Productivity". McKinsey Global Institute, pp. 1-137 (2011)

[21] A European Commission survey has recently found that data protection remains a major concern for EU citizens (July 2015), http://www.technologyslegaledge.com/2015/07/07/europe-european-commission-survey-finds-that-data-protection-remains-a-major-concern-for-eu-citizens/, accessed December 2015.

[22] https://www.enisa.europa.eu/activities/Resilience-and-CIIP/cloud-computing/big-data-security, accessed January 2016.

detailed into five layers: “Data sources”, “Integration process”, “Data storage”, “Analytics and computing models“, “Presentation”.

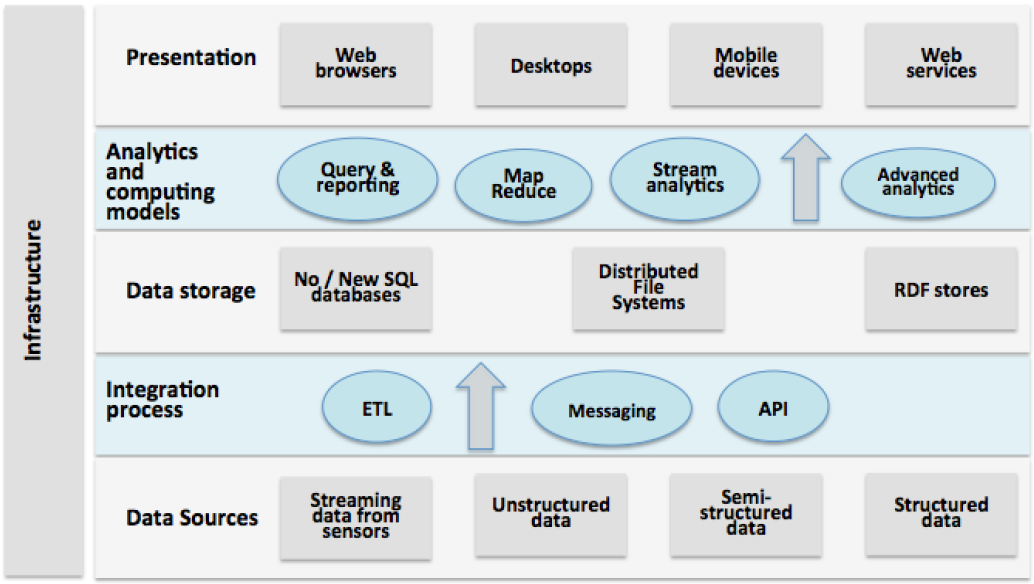


Figure 2–1 Layered architecture of Big Data systems

The function of each layer is as follows:

The “Data sources” layer consists of disparate data sources, ranging from sensor streaming data, to structured information such as relational databases, and to any sort of unstructured and semi-structured data.

The “Integration process” layer is concerned with acquiring data and integrating the datasets into a unified form with the necessary data pre-processing operations.

The “Data storage” layer consists of a pool of resources such as distributed file systems, RDF stores, NoSQL and NewSQL databases, which are suitable for the persistent storage of a large number of datasets.

The “Analytics and computing models” layer encapsulates various data tools, such as Map Reduce, which run over storage resources and include the data management and the programming model[23]. The “Presentation” layer enables the visualisation technologies.

Cloud computing can be deployed as the infrastructure layer for Big Data systems to meet some infrastructure requirements, such as cost-effectiveness, elasticity, and the ability to scale up or down[24].

[23] Map Reduce is a programming framework that has achieved great success in processing group-aggregation tasks. Hadoop is an open source tool that implements Map Reduce framework. Hadoop integrates data storage, data processing, system management, and other modules to form a system-level solution, which is becoming the mainstay in handling Big Data challenges, https://hadoop.apache.org/, accessed December 2015.

[24] Hu, H. (School of Computing, National University of Singapore), Wen, Y., Chua, T.-S. & Li, X. “Toward Scalable Systems for Big Data Analytics: A Technology Tutorial”. IEEE Access 2, 652–687 (2014).

3. Big Data assets

Assets can be abstract assets (like processes or reputation), virtual assets (for instance, data), physical assets (cables, a piece of equipment), human resources, money". An item of our taxonomy is either a description of data itself, or describes assets that generate, process, store or transmit data chunks and, as such, is exposed to cyber-security threats. For information security considerations, this study focuses on assets that are related mainly to information and communication technology (ICT) under the scope of Big Data.

A major source of information for this study is the work made by the NIST Big Data Public Working Group (NBD-PWG)[25], which is developing consensus on important and fundamental questions related to Big Data. They have produced two draft Volumes (Volume 1 about Definitions and Volume 2 about Taxonomy). Another source of information is the report "Big Data Taxonomy"[26], issued by Cloud Security Alliance (CSA) Big Data Working Group in September 2014. In that document, CSA proposes a six-dimensional taxonomy for Big Data, pivoted around the nature of the data to be analysed. The objective is to help "navigate the myriad choices in compute and storage infrastructures as well as data analytics techniques" and the proposed structure is mainly intended as a high-level taxonomy for decision makers.

Specifically, most of the terminology used in this report for high level asset types (*Data*, *Infrastructure*, *Analytics*, and *Security and Privacy techniques*) comes, with some small modifications, from the CSA taxonomy; where our term *Infrastructure* also comprises of the other two CSA main categories; viz. *Compute Infrastructure* and *Storage Infrastructure*. Another high-level type, *Roles,* comprises human resources and other related assets, as in previous ENISA thematic studies[27].

3.1 Big Data asset taxonomy

Error! Reference source not found. gives an overview of the Big Data assets structure into relevant categories according to their use. The full list of identified the Big Data assets is given in Annex A.

[25] NIST Special Publication 1500-1, "DRAFT NIST Big Data Interoperability Framework: Volume 1, Definitions", by NIST Big Data Public Working Group Definitions and Taxonomies Subgroup (Draft Version 1 April 6, 2015) and NIST Special Publication 1500-2 "DRAFT NIST Big Data Interoperability Framework: Volume 2, Big Data Taxonomies" by NIST Big Data Public Working Group Definitions and Taxonomies Subgroup (Draft Version 1 April 6, 2015), http://bigdatawg.nist.gov/V1_output_docs.php, accessed December 2015.

[26] Cloud Security Alliance BIG DATA WORKING GROUP, "Big Data Taxonomy" September 2014, https://cloudsecurityalliance.org/download/big-data-taxonomy/, accessed December 2015.

[27] For example: "Smart Grid Threat Landscape" (Dec 2013), "Threat Landscape and Good Practice Guide for Internet Infrastructure", "Threat Landscape and Good Practice Guide for Smart Home and Converged Media", https://www.enisa.europa.eu/activities/risk-management/evolving-threat-environment/enisa-thematic-landscapes, accessed December 2015.

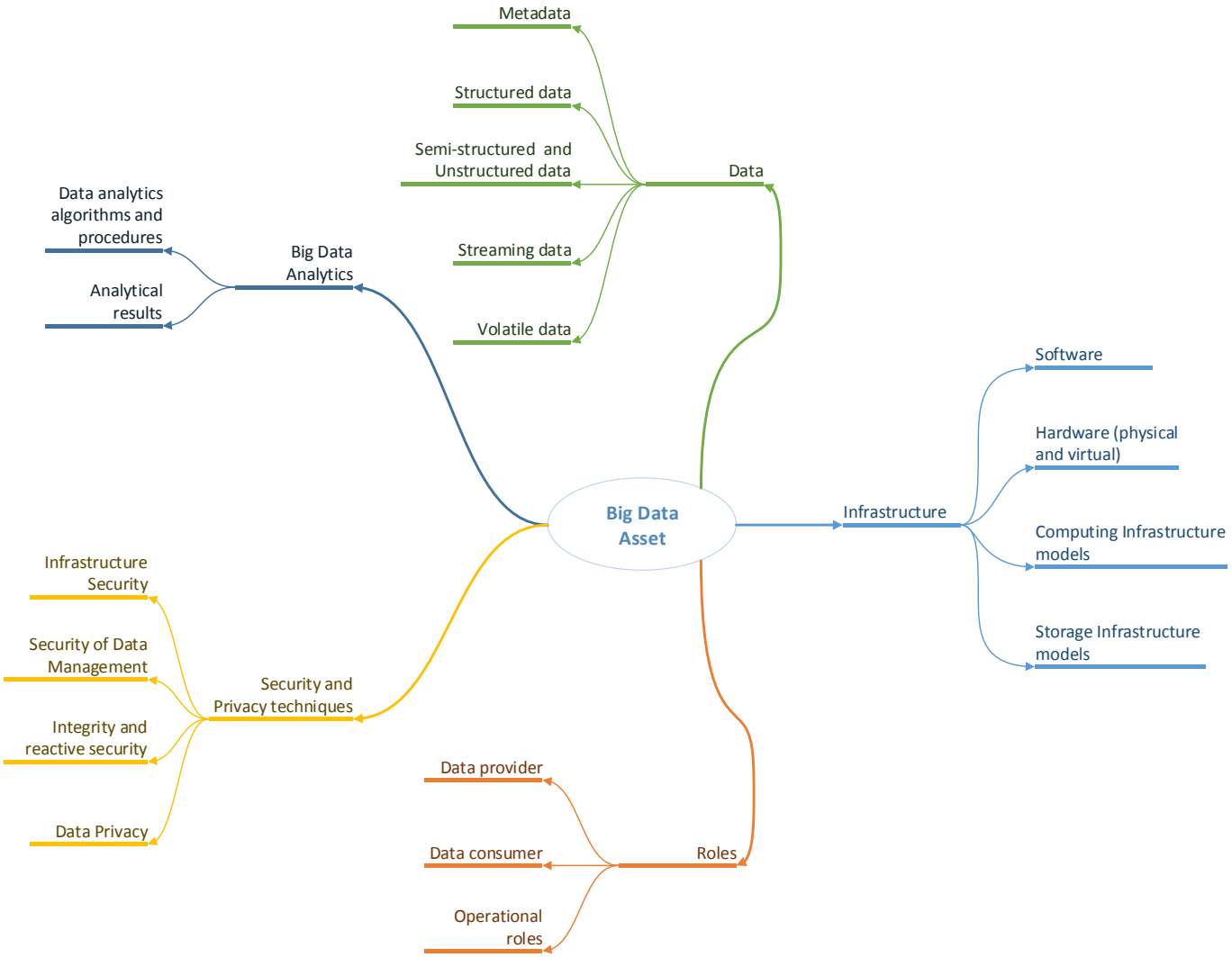


Figure 3–1 Big Data asset taxonomy (asset groups and asset types only)

3.2 Big Data asset categories

With the following list we attempt to identify some of the known Big Data valuable assets in a hierarchical manner. The first and second level category items (asset group and asset type) can be thought of as intuitively clear, but we give a brief description of them nevertheless. Numbers in brackets refer to **Error! Reference source not found.**. The full taxonomy, with further levels in the taxonomy such assets and asset details, is presented in Annex A[28].

Data – This is the core category of the Big Data taxonomy and includes:

Metadata, i.e. schemas, indexes, data dictionaries and stream grammars' data (which often but not necessarily come together with stream data).

Structured data, i.e. database records structured according to a data model, as for example a relational or hierarchical schema; structured identification data, as for example users' profiles and preferences; linked open data; inferences and re-linking data structured according to standard formats.

Semi-structured and unstructured data, for example logs, messages and web (un)formatted data (Web and Wiki pages, e-mail messages, SMSs, tweets, posts, blogs, etc.), files and documents (e.g. PDF files and Office suite data in Repositories and File Servers), multimedia data (photos, videos, maps, etc.), and other non-textual material besides multi-media (medical data, bio-science data and raw satellite data before radiometric/geometric processing, etc.).

Streaming data, i.e. single-medium streaming (for example in-motion sensor data) and multimedia streaming (remote sensing data streams, etc.).

Volatile data, i.e. data that are either in motion or temporarily stored, as, for example, network routing data or data in devices' random access memory.

Infrastructure – The term infrastructure comprises software, hardware resources denoting both physical and virtualized devices, the basic computing infrastructure with its batch and streaming processes and the storage infrastructure with all sort of database management systems, ranging from old-style relational databases to NoSQL or NewSQL, as well as Semantic Web tools. Specifically, the Infrastructure first level category includes:

Software, including operating systems, device drivers, firmware, server-side software packages (as Web and Application Server software) and applications. Applications sub-category includes software implementation as back-end services and all sorts of functionalities that utilize other assets in order to fulfil a defined task, such as for example asset management tools, requirements gathering applications, billing services and tools to monitor performances and SLAs.

Hardware (physical and virtual), i.e. servers (physical devices and hardware nodes, all the virtualized hardware, including virtual Data Centres with their management consoles and virtual machines, as well as the physical hardware supporting their provisioning), clients, network devices (for example, physical switches, virtual switches and virtual distributed switches, etc.), media and storage devices (the various

[28] In the proposed taxonomy an asset could be a member of more than one category (e.g., streaming data could be both structured and unstructured data). This choice is due to the fact that this guide is mainly aimed at identifying threats or groups of threats that could affect very broad, and, in some cases, even overlapping asset categories. Also, this choice may allow a better correlation between threats.

types of disk storage, etc.), data gathering devices (sensors, remote platforms as airborne platforms or drones, etc.), Human Interface Devices (HID) and mobile devices.

Computing Infrastructure Models, this category includes paradigms of abstract processing architectures, on whether the processing can be done in batch mode, for example MapReduce; on real-time/near real-time streaming data, as for example Sketch or Hash-based models; or follow a unified approach supporting both, as for example Cloud Dataflow.

Storage Infrastructure Models, this category includes paradigms of abstract storage architectures, including Big Files and triples-based models.

Big Data Analytics – This category includes models which define protocols and algorithms for Big Data analysis, like procedures, models, algorithms definitions down to the source code, and analysis' results. The category includes:

Data analytics algorithms and procedures, which include algorithm source code with their set-up parameters, configuration and thresholds, metrics, the model definitions, advanced techniques that streamline the data preparation stage of the analytical process.

Analytical results, either in textual or in graphical mode (e.g. spatial layouts, abstract, interactive and real time visualizations).

Security and Privacy techniques – This category name includes the term "techniques" to remark that the security-related assets it includes are the ones of interest to attackers and therefore more subject to unauthorized disclosure and leakage, as for example security best practice documents, cryptography algorithms and methods, information about the access control model used, etc. The category includes the following sub-categories:

Infrastructure Security, i.e. the first aspect of a Big Data ecosystem security, which deals with how to secure the distributed computation systems and the data stores, with security Best Practices and policy set-ups.

Data Management, i.e. documents and techniques about how to secure Data Storage and Logs, and documentation about granular audits and data life cycle (Data provenance).

Integrity and reactive security, which deals with all the practices, techniques, and documents related to End Point validation and filtering and the monitoring of real-time security, including incident handling and information forensics.

Data Privacy, i.e. all the techniques put in place to protect privacy as it is requested by law, for example cryptographic methods and access control.

Roles - This terminology for this category was introduced by the NIST Big Data Public Working Group[29] and includes:

Data provider, such as enterprises, organizations, public agencies, academia, network operators and end-users.

[29] For a more detailed description of the Human Resources assets see Annex A.

Data consumer, partly overlapping the previous category, but from a different scope, and including enterprises, organizations, public agencies, academia and end-users.

Operational roles, i.e. system orchestrators (business leader, data scientists, architects, etc.), Big Data application providers (application and platform specialists), Big Data framework providers (Cloud provider personnel), security and privacy specialists, technical management (in-house staff, etc.).

We remark that leaving the taxonomy unbalanced (some sub trees, like those rooted in Data and Infrastructure are deeper than others) is a deliberate choice. Indeed some leaf subcategories of our taxonomy, such as *Models definitions*, could be used to integrate external taxonomies designed for different reasons, such as data science ones[30].

Another remark is that most of the categories and sub-categories could be related to data, rather than Big Data. For example, relational databases are a very typical and common resource in every enterprise infrastructure, not necessarily storing big data volumes. Even when relational databases have big volume size, they are often manageable through traditional hardware clusters, appliances and software tools. Another example is applications' random-access memory (featured in volatile data category), i.e. the data that is temporarily in memory due to processing operations. This memory is often (though not invariably, as witnessed by the success of in-memory processing systems) not large, compared to massive data sizes of in-memory databases.

Nevertheless, we included these assets in our taxonomy for completeness of information. Data stored in relational databases, often very valuable for data owners, might be used in some cases as data source for analytics, while leakage of RAM content could compromise login credentials and cryptographic keys, paving the way to dangerous attacks to Big Data.

The presented asset taxonomy should only be considered as a snapshot of the complex range of Big Data assets and could as such not be exhaustive.

[30] See for instance https://www.thoughtworks.com/insights/blog/data-science-ontology, accessed December 2015.

4. Big Data threats

4.1 ENISA threat taxonomy

In this section, we introduce the major characteristics of the ENISA threat taxonomy. The ENISA taxonomy is a comprehensive one, with a special focus on cyber-security threats; i.e., threats applying to information and communication technology assets. Additional non-ICT-stemming threats have been considered to cover threats to physical assets and also both natural disasters [not directly triggered by humans] and environmental disasters directly caused by human.

The threat taxonomy has been developed by the ENISA Threat Landscape (ETL) Group and is a consolidation of threats previously considered in other thematic reports[31] and extensive research. The taxonomy includes threats applicable to the Big Data assets and only these are depicted in figure 4-1. In the following subsection, threats specific to Big Data that were identified through extensive literature that have been assigned to the relevant categories defined in ENISA's Threat Taxonomy are mapped to the previously discussed Big Data Asset Taxonomy.

[31] Smart Grid Threat Landscape, Threat Landscape and Good Practice Guide for Internet Infrastructure, Threat Landscape and Good Practice Guide for Smart Home and Converged Media

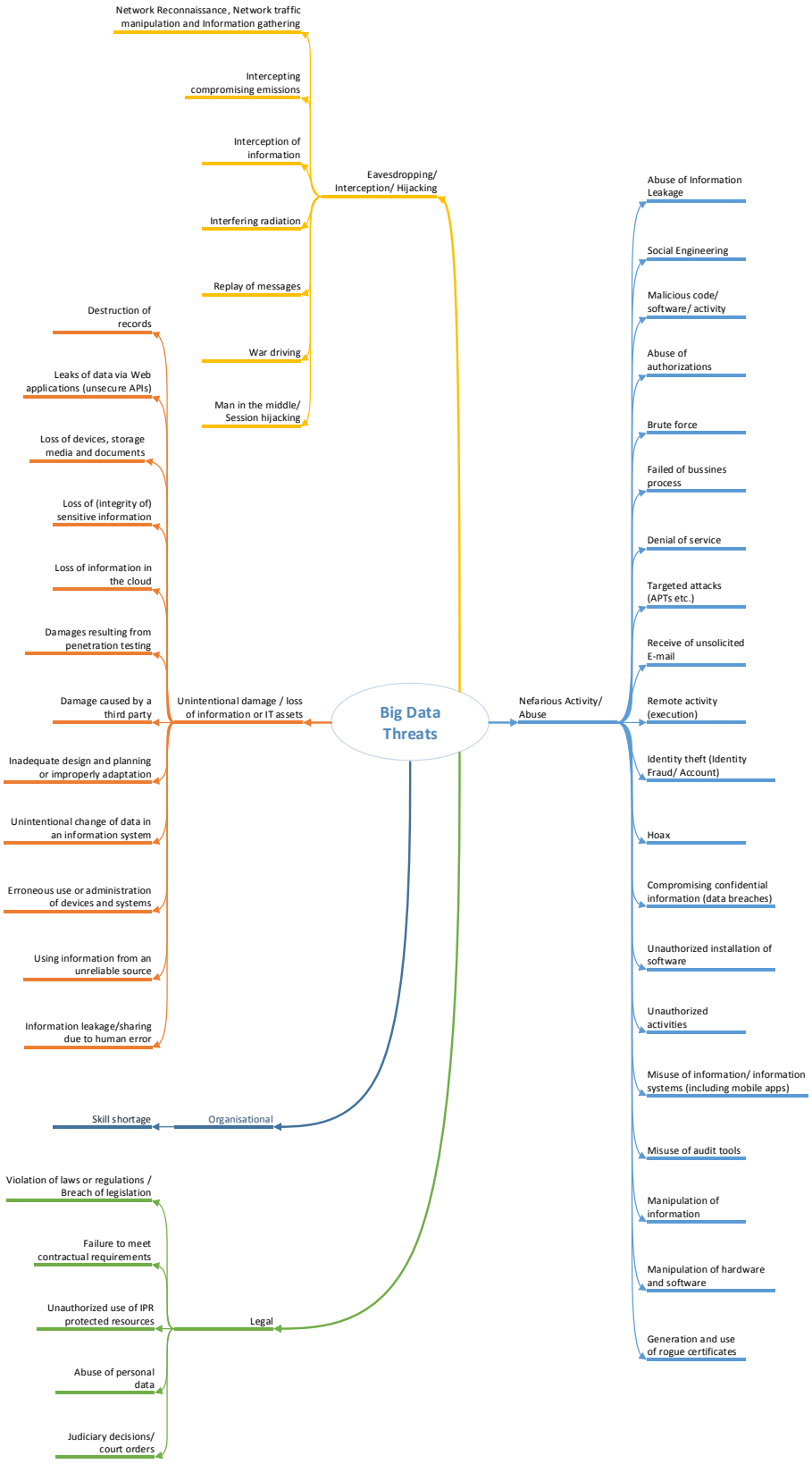


Figure 4–1 Threat taxonomy applicable to Big Data assets

4.2 Mapping threats to Big Data assets

In this section, we discuss the threats that can be mapped to the Big Data asset taxonomy presented in the previous chapter. This analysis is based on an extensive review of actual threat incidents and attacks to Big Data presented in articles, technical blogs, conference papers, as well as online surveys for gathering supplemental information. Our review was driven by the ENISA generic threat taxonomy presented in the previous section.

In general terms, threats, such as network outage or malfunctions of the supporting infrastructure, may heavily affect Big Data. In fact, since a Big Data has millions of pieces of data and each piece may be located in a separate physical location, this architecture leads to a heavier reliance on the interconnections between servers. Past ENISA thematic reports have dealt in depth with threats such as outages and malfunctions, which affect network communication links[32]. For this reason, in this report, we don't take these threats into account. Also, we chose not to dwell on physical attacks (deliberate and intentional), natural and environmental disasters, and failures / malfunction (e.g. malfunction of the ICT supporting infrastructure), since their effects are strongly mitigated by the intrinsic redundancy of Big Data, though Big Data owners deploying their systems in private clouds or other on-premise infrastructure should take these attacks under serious consideration [33].

In general, a threat is "any circumstance or event with the potential to adversely impact an asset through unauthorized access, destruction, disclosure, modification of data, and/or denial of service"[34]. Given the definition we gave of Big Data (*Volume*, *Velocity*, *Variety*, *Veracity*, *Variability* and *Value*), a threat to a Big Data asset can be considered as any circumstance or event that affects, often simultaneously, big volumes of data and/or data in various sources and of various types and/or data of great value.

We also identify two different sub-categories of threats: "Big Data Breach" and "Big Data Leak"[35], orthogonal to the used threat taxonomy. A breach occurs when "a digital information asset is stolen by attackers by breaking into the ICT systems or networks where it is held/transported"[36]. We can define "Big Data Breach" as the theft of a Big Data asset executed by breaking into the ICT infrastructure. A Big Data Leak on the other hand, can be defined as the (total or partial) disclosure of a Big Data asset at a certain stage of its lifecycle. A Big Data Leak can happen for example in inadequate design, improper software adaptation or when a business process fails. In terms of the attacker model, a Big Data Breach requires pro-active hostile behaviour (the break-in), while a Big Data Leak can be exploited even by honest-but-curious attackers.

---

[32] ENISA, "Threat Landscape and Good Practice Guide for Internet Infrastructure", January 2015

[33] We must note that some basic, trial or special purpose Big Data installations might not, fully or partially, rely on [cloud] redundancy. For example Big Data installed in a private cloud environment, with no replication services enabled, has no disaster recovery option in case of a natural disaster. Being limited to a private local cloud environment, these installations could, in principle, be subject to physical attacks and natural and environmental disasters, such as earthquakes, floods, landslides, tsunamis, fire, pollution, dust, thunder stroke, and other major events for the environment. However, enabling private cloud replication services between different physical locations mitigates this risk.

[34] See glossary in https://www.enisa.europa.eu/activities/risk-management/current-risk/risk-management-inventory/glossary, accessed December 2015.

[35] E. Damiani, "Toward Big Data Leak Analysis". Proceedings of Privacy and Security of Big Data Workshop (PSBD 2015), IEEE Big Data Conference, San Jose, CA, 1-3 November 2015

[36] See ISO 15408 model.

4.2.1 Threat Group: Unintentional damage / loss of information or IT assets

This group includes Information leakage or sharing due to human errors, unintentional intervention or erroneous use of administration of systems (misconfiguration), loss of devices.

Threat: Information leakage/sharing due to human error

Accidental threats are those not intentionally posed by humans. They are due to misconfiguration, skill-based slips and clerical errors (for example pressing the wrong button), misapplication of valid rules (poor patch management, use of default user names and passwords or easy-to-guess passwords), and knowledge-based mistakes (software upgrades and crashes, integration problems, procedural flaws)[37] [38].

Information leakage due to misconfiguration can be a common problem: according to a recent study[39], erroneous system administration setups led to numerous weaknesses in four different Big Data technologies; viz. Redis, MongoDB, Memcache and ElasticSearch. According to the same study most of these new products "*are not meant to be exposed to the Internet. […] These technologies' default settings tend to have no configuration for authentication, encryption, authorization or any other type of security controls that we take for granted. Some of them don't even have a built-in access control.*"

Furthermore, in the past, there have been reported incidents of inappropriate sharing of files containing possible sensitive and confidential information, which affected even very popular online services like Dropbox[40]. This is also confirmed by many surveys[41].

The assets targeted by these threats include asset group "**Data**", and asset "**Applications and Back-end services**" (such as for example "**Billing services**").

---

[37] See the human error taxonomy in information systems in Im and Richard L. Baskerville (Georgia State University), "A Longitudinal Study of Information System Threat Categories: The Enduring Problem of Human Error". ACM SIGMIS (2005).

[38] According to "IBM Security Services 2014 Cyber Security Intelligence Index" over 95 percent of all incidents investigated recognize "human error" and the most prevalent contributing human error? "Double clicking" on an infected attachment or unsafe URL.

[39] BinaryEdge, a Switzerland-based security-engineering firm, probed four common Big Data technologies, such as Redis, MongoDB, Memcache, and Elasticsearch, and found various configuration problems. For example, the company found that dozens of thousands of instances of NoSQL databases were accessible without any authentication required. See http://blog.binaryedge.io/2015/08/10/data-technologies-and-security-part-1/, accessed December 2015.

[40] Techcrunch, a popular online publisher of technology industry news, reported that at Dropbox, for a brief period of time, the service allowed users to log into accounts using any password. In other words, people could log into someone's account simply by typing in their email address. See http://techcrunch.com/2011/06/20/dropbox-security-bug-made-passwords-optional-for-four-hours/, accessed December 2015.

[41] The vast majority (more than 80%) of participants in EMA Research's 2015 State of File Collaboration Security report, sponsored by FinalCode, admitted that there have been data leakage incidents in their organizations. See http://www.finalcode.com/en/how-it-works/resources/ema-report/, accessed December 2015.

Threat: Leaks of data via Web applications (unsecure APIs)

Various sources claim that Big Data is often built with little security[42] [43]. New software components are usually provided with service-level authorization, but few utilities are available to protect core features and application interfaces (APIs). Since Big Data applications are built on web services models, APIs may be vulnerable to well-known attacks, such as the Open Web Application Security Project (OWASP) Top Ten list[44], with few facilities for countering common web threats.

The security software vendor Computer Associate (CA) [45] and other sources[46] report data breaches, due to insecure APIs, in many industries, especially in social networks, mobile photo-sharing and video-sharing services, as Facebook, Yahoo and Snapchat.

For example, a threat of this category may consist in injection attacks to Semantic Web technologies through SPARQL code injection[47]. Security flaws are rather common in new Big Data languages like SPARQL, RDQL (both are read-only query languages) and SPARUL (or SPARQL/Update, which has modification capabilities). The use of these new query languages introduces vulnerabilities already found in a bad use of old-style query languages, since attacks like SQL, LDAP and XPath injection are already well known and still dangerous[48]. Libraries of these new languages provide tools to validate user input and minimize the risk. However, "*main ontology query language libraries still do not provide any mechanism to avoid code injection*" and without these mechanisms, attackers' arsenal might get enhanced with SPARQL, RDQL and SPARQL injections[49]. Other new Big Data software products, as for example Hive, MongoDB and CouchDB, also suffer from traditional threats such as code execution and remote SQL injection[50].

---

[42] Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments, released by the security company Securosis, in October 2012, https://securosis.com/assets/library/reports/SecuringBigData_FINAL.pdf, accessed December 2015.

[43] Eduardo B. Fernandez (Department of Computing and Electronic Engineering and Computer Science, Florida Atlantic University), "Security in Data Intensive Computing Systems in Handbook of Data Intensive Computing". Springer (2011), http://link.springer.com/chapter/10.1007/978-1-4614-1415-5_16, accessed December 2015.

[44] Many common vulnerability exposures for Big Data components, such as Hadoop, are reported in specialized Websites, see for example https://cve.mitre.org and https://www.cvedetails.com, accessed December 2015.

[45] Jaime Ryan (CA, Sr. Director) and Tyson Whitten (CA, Director of API Management) in "Takeaways from API Security Breaches" presentation and webinar (2015) reported breaches, due to unsecure APIs, for Yahoo, Snapchat and other companies, see http://transform.ca.com/API-security-breaches.html?source=AAblog, accessed December 2015.

[46] See security issues for the Graph Facebook API library reported by Websegura technical blog, http://www.websegura.net/advisories/facebook-rfd-and-open-file-upload/, accessed December 2015.

[47] See http://www.morelab.deusto.es/code_injection/ and the following publication: Pablo Orduña, Aitor Almeida, Unai Aguilera, Xabier Laiseca, Diego López-de-Ipiña, Aitor Gómez-Goiri, "Identifying'Identifying Security Issues in the Semantic Web: Injection attacks in the Semantic Query Languages", [*VI Jornadas Científico-Técnicas en Servicios Web y SOA (JSWEB 2010*p.)], Valencia, Spain. September 2010, pp. 43 - 50. ISBN: 978-84-92812-59-2.

[48] In October 2015, presumably, an SQL injection was used to attack the servers of British telecommunications company Talk Talk's, endangering the personal details of up to four million customers. See http://www.mobilenewscwp.co.uk/2015/10/23/talktalk-hacking-scandal-expert-reaction/, accessed December 2015.

[49] Ben Mustapha et al., "Enhancing semantic search using case-based modular ontology". in Proceeding of the 2010 ACM Symposium on Applied Computing.

[50] For example Hive version 2.0 suffers from cross site scripting, code execution, and remote SQL injection vulnerabilities, see https://packetstormsecurity.com/files/132136/Hive-2.0-RC2-XSS-Code-Execution-SQL-Injection.html. MongoDB suffers njection attacks, see https://www.idontplaydarts.com/2011/02/mongodb-null-byte-injection-attacks/. See also some other vendor-specific threats in presentation

The assets targeted by these threats belong to group "**Data**" and asset type "**Storage Infrastructure models**" (such as "**Database management systems (DBS)"** and "**Semantic Web tools**")

Threat: Inadequate design and planning or incorrect adaptation

Techniques for improving Big Data analytics performance and the fusion of heterogeneous data sources increase the hidden redundancy of data representation, generating ill-protected copies. This challenges traditional techniques to protect confidentiality[51] and the effect of redundancy must be taken into account. As already stated, Big Data redundancy can be seen as a threat mitigation technique for physical attacks, disasters and outages[52], however in some cases it signals a system weakness, being a risk booster for Big Data leaks. In other words, if our Big Data storage replicates data records ten times and distributes the copies to ten storage nodes for some reason (e.g., to speed up the analytics pipeline), the ten nodes may end up with different levels of security robustness (e.g., different security software versions) and this will increase the probability of data disclosure and data leaks. This can be considered a specific weakness of Big Data designs.

On the other hand we can also note that even the redundancy and the replication that are necessary features to enhance Big Data functionality, are not always a failsafe against data loss. For example Hadoop, the well-known framework for Big Data processing, replicates data three times by default, since this protects against inevitable failures of commodity hardware. However, a corrupted application could destroy all data replications[53]. Also, recent studies put forward the idea that Hadoop redundancy could even be a non-linear risk booster for Big Data leakages[54].

Even the design of the Hadoop Distributed File System (HDFS) signals problems as reported by literature[55]. HDFS is the basis of many Big Data large-scale storage systems and is used by social networks. HDFS clients perform file system metadata operations through a single server known as the Namenode, and send and retrieve file system data by communicating with a pool of nodes. The loss of a single node should never be fatal, but the loss of the Namenode cannot be tolerated[56]. Big social networks, such as Facebook, suffered this problem and took countermeasures against the threat[57] (Hadoop installed at Facebook includes one of the largest single HDFS cluster, more than 100 PB physical disk space in a single HDFS file system).

---

https://www.defcon.org/images/defcon-21/dc-21-presentations/Chow/DEFCON-21-Chow-Abusing-NoSQL-Databases.pdf, accessed December 2015.

[51] E. Damiani, "Toward Big Data Risk Analysis", Keynote Speech at the 2nd International Workshop on Privacy and Security of Big Data (PSBD 2015)

[52] Physical attacks, Disasters and Outages are respectively described as threat group in the generic ENISA threat taxonomy.

[53] See http://www.smartdatacollective.com/michelenemschoff/193731/how-your-hadoop-distribution-could-lose-your-data-forever, accessed December 2015.

[54] E. Damiani, "Toward Big Data Leak Analysis", *Proceedings of the Privacy and Security of Big Data Workshop (PSBD 2015)*, IEEE Big Data Conference, San Jose, CA, 1-3 November 2015

[55] Aditham, Ranganathan (Dept of Computer Science and Engineering, University of South Florida, Tampa, USA), "A Novel Framework for Mitigating Insider Attacks in Big Data Systems". 2015 IEEE International Conference on Big Data

[56] All metadata operations go through the Namenode. If the Namenode is unavailable, no clients can read from or write to HDFS, and users and applications that depend on HDFS will not be able to function properly. Recent versions of Hadoop have introduced other components for resource management to deal with this issue.

[57] See "Notes by Facebook engineering" in https://www.facebook.com/notes/facebook-engineering/under-the-hood-hadoop-distributed-filesystem-reliability-with-namenode-and-avata/10150888759153920. Facebook contributed to a working solution to address the architectural shortcomings of the single Namenode failover, called Avatarnode. This

One more threat related to the design is the lack of scalability of some tools. For example NIST reports that original digital rights management (DRM) techniques were not built to scale and to meet demands for the forecasted use of the data and "*DRM can fail to operate in environments with Big Data characteristics—especially velocity and aggregated volume*"[58] [59].

The assets that are targeted by these threats belong to asset groups "**Data**" and "**Big Data analytics**", and to asset types "**Software**", "**Computing Infrastructure models**" and "**Storage Infrastructure models**".

4.2.2 Threat Group: Eavesdropping, Interception and Hijacking

This group includes threats that rely on alteration/manipulation of the communications between two parties. These attacks do not require installing additional tools or software on the victims' infrastructure.

Threat: Interception of information

A common issue that affects any ICT infrastructure is when offenders can intercept communications between nodes by targeting the communication links. Various sources claim that inter-node communication with new Big Data tools is often unsecured[60], that it is not difficult to hijack a user session or gain unauthorized access to services in social networks as Facebook and Twitter[61], and that there is evidence of flaws in communication protocols[62].

Big Data software distributions (for example Hadoop, Cassandra, MongoDB[63], Couchbase) rarely have the protocols that ensure data confidentiality and integrity between communicating applications (e.g., TLS and SSL) enabled by default or configured properly (e.g., changing default passwords).

The assets targeted by this threat belong to asset groups "**Data**" and "**Roles**", and to asset "**Applications and Back-end services**".

4.2.3 Threat Group: Nefarious Activity/Abuse

This group includes threats coming from nefarious activities. Unlike the previous group, these threats (often) require the attacker to perform some actions altering the victims' ICT infrastructure; usually with the use of specific tools and software.

---

is open source module offering hot failover and failback and is now in production at Facebook running the largest Hadoop Data Warehouse cluster (100 PB physical disk space in a single HDFS file system).

[58] See NIST Special Publication 1500-4. Use case: consumer digital media (examples: Netflix, iTunes, and others).

[59] Xiao Zhang, "A Survey of Digital Rights Management Technologies", see http://www.cse.wustl.edu/~jain/cse571-11/ftp/drm.pdf, accessed December 2015.

[60] *Securing Big Data: Security Recommendations for Hadoop and NoSQL Environments*, Released by security company Securosis, L.L.C., October 12, 2012.

[61] See for example "How to prevent a session hijacking attack" in Facebook and Twitter, http://searchmidmarketsecurity.techtarget.com/tip/Defending-against-Firesheep-How-to-prevent-a-session-hijacking-attack, accessed December 2015.

[62] See for example an attack against confidentiality of data in transit across untrusted networks in http://www.isg.rhul.ac.uk/tls/Lucky13.html, accessed December 2015.

[63] See, for example, MongoDB documentation about and mistakes that can compromise the database (TTL configuration errors and others), http://blog.mongodb.org/post/87691901392/mongodb-security-part-ii-10-mistakes-that-can, accessed December 2015.

Threat: Identity fraud

Big Data systems store and manage credentials for accessing personal data and financial accounts with information such as credit card numbers and payment and billing details, which are targets for cyber criminals. Big Data systems also store profiling data that can describe user behaviour, preferences, habits, travel, media consumption at a high degree of detail, and may help attackers in more elaborate forms of impersonation fraud, creating big opportunities for identity thieves[64].

Since most Big Data systems are built on top of cloud infrastructure, a threat to users' identity is, for example, when the control of a system interface, in either a Big Data system based on a large public cloud or in a widely used private cloud, gets lost[65]. A successful attack on a console grants the attacker complete power over the victim's account, including all the stored data. The control interfaces could be initially compromised via novel signature wrapping and advanced XSS techniques, then privilege escalation may lead to identity fraud[66]. While in traditional information systems the loss of control of a console interface could cause limited information leakage, in Big Data the effect is amplified and the impact is more severe.

Social engineering is not a new issue, but as social networking becomes important both for home users and businesses, attacks often involve social engineering. Attackers have been abusing social networks since they first came online. For example, XSS vulnerabilities on Twitter have been used to push malicious and fake tweets, while Internet malware has emerged on Facebook as a means of promoting malicious profiles[67].

The assets targeted by these threats are "**Personal identifiable information**", "**Applications and Back end services**" (such as, for example, "**Billing services**") and "**Servers**".

Threat: Denial of service

Big Data components can be threaten by traditional denial of service (DoS) and distributed denial of service (DDoS) attacks. For example, such attacks may remove Big Data components from the network and then exploit its vulnerabilities or an attacker could exhaust the limited resources in a Hadoop cluster, leading to a significant decrease of system performance and causing the loss of the targeted resource to other cloud users[68]. But, at the same time, countering mechanisms have been developed for/using Big

---

[64] Big data creates big opportunities for identity thieves: see http://www.c4isrnet.com/story/military-tech/it/2015/01/19/big-data-identity-theft/22004695/, accessed December 2015.

[65] See: J. Somorovsky et al., "All your clouds belong to us: security analysis of cloud management interfaces", in *Proceedings of the 3rd ACM workshop on Cloud computing security workshop* (http://dl.acm.org/citation.cfm?id=2046664) for attacks to Amazon and Eucalyptus. The paper provides a security analysis pertaining to the control interfaces of large public Cloud services (Amazon EC2 and S3) and private Cloud software (Eucalyptus).

[66] See http://www.zdnet.com/article/us-cert-warns-of-guest-to-host-vm-escape-vulnerability/. The article describes a vulnerability, which affects 64-bit operating systems and virtualization software running on Intel CPU hardware, and exposes users to local privilege escalation attack or a guest-to-host virtual machine escape.

[67] See Nine Threats Targeting Facebook Users in http://www.itbusinessedge.com/slideshows/show.aspx?c=90875, accessed December 2015.

[68] Jingwei Huang, David M. Nicol, and Roy H. Campbell (Information Trust Institute, University of Illinois at Urbana-Champaign, Illinois), "Denial-of-Service Threat to Hadoop/YARN Clusters with Multi-Tenancy". IEEE International Congress on Big Data (2014).

Data systems. For example administrators of Hadoop infrastructure can deploy specialized components to track DDOS attacks[69].

In the past this kind of attacks has led to some service outages for Amazon distributed storage, through elevated levels of authenticated requests and account validation[70]. Furthermore, as already stated, also specific attacks against social networks such as Facebook have been mounted, exploiting some weaknesses of the Hadoop Distributed File system, for example the Namenode single server[71].

Assets targeted by this threat include the asset "**Servers**" (viz. **Virtualized Data Centre**", "**Physical Machine**" and "**Virtual Machine**") and the asset "**Network**".

Threat: Malicious code / software / activity

These very generic threats affect almost all the ICT components of an infrastructure. Examples of these threats are: *i)* ***exploit kits***, which allow virus and malware infections, *ii)* ***worms***, which may be distributed by using the network to send copies to other nodes, *iii)* ***Trojans***, which are pieces of malware that facilitate unauthorized access to a computer system, *iv)* ***backdoors and trapdoors***, which are undocumented entry points into a computer program, generally inserted by a programmer to allow remote access to the program, *v)* ***service spoofing***, which is an attack in which the adversary successfully masquerades as another by falsifying data and thereby gaining an illegitimate advantage[72], *vi)* ***web application attacks and injection attacks*** through code injection –examples of exploiting this threats to mount more elaborate attacks have already been discussed –.

After the deployment of the code, the attacker may manipulate infected devices. In Big Data, malware infected nodes may send targeted commands to other servers and disturb or manipulate their operations, worms may replicate themselves sending copies to other nodes and affect the behaviour of all components connected to the network. There is always the possibility that vendors of Big Data tools, or somebody else in the software chain, may have installed firmware with backdoors or some hidden functionality to facilitate access to the devices, in particular in the context of very new technologies such as NoSQL and NewSQL[73].

An example of hacking Big Data through a malicious code attack is reported in literature[74] as faulty results of the Hadoop logging data system. System administrators use Hadoop server logs to identify potential attacks. A demo of this hack requires that a service, called Flume, streams logs into a SQL based Hadoop data store (Hcatalog). In this scenario, an attacker runs a malicious script and alters the results by

---

[69] A Hadoop service called Flume can be used for streaming log data transfer into Hcatalog (a SQL based Hadoop data store). An example of system log monitoring is given in a tutorial by Hortonworks, where DDOS attacks are being tracked down by system admins. See http://hortonworks.com/hadoop-tutorial/how-to-refine-and-visualize-server-log-data/, accessed December 2015.

[70] See ZDnet bog in http://www.zdnet.com/article/amazon-explains-its-s3-outage/, accessed December 2015.

[71] See "Notes by Facebook engineering" in https://www.facebook.com/notes/facebook-engineering/under-the-hood-hadoop-distributed-filesystem-reliability-with-namenode-and-avata/10150888759153920, accessed December 2015.

[72] A concrete example of spoofing is ARP spoofing in the MAC layer: the management frames are not authenticated in 802.11. Every frame has a source address. The attackers take advantage of the spoofed frame to redirect the traffic and corrupt the ARP tables.

[73] Eweka Raphael Osawaru, Riyaz Ahamed, "A Highlight of Security Challenges in Big Data". International Journal of Information Systems and Engineering (online), Volume 2, Issue 1 (April 2014).

[74] Aditham, Ranganathan (Dept of Computer Science and Engineering, University of South Florida, Tampa, USA), "A Novel Framework for Mitigating Insider Attacks in Big Data Systems". 2015 IEEE International Conference on Big Data

modifying the log data before Flume can stream them into Hcatalog[75]. The logs can be corrupted even when Hadoop services seem to be working as expected.

Malicious software can be a threat also in distributed programming frameworks, which use parallel computation, and may have untrusted components. For example, MapReduce computational framework splits the input file into multiple chunks: in the first phase a mapper reads the data, performs computation, and outputs key/value pairs. In the second phase, a reducer works on these pairs and outputs the result. A key issue is how to secure the mappers[76], since untrusted mappers alter results. With large data sets, it becomes difficult to identify malicious mappers.

The assets targeted by this attack include "**Database management systems (DBMS**)" (such as the traditional "**Relational SQL**" databases, and the Big Data new tools "**NoSQL**" and "**NewSQL**"), and asset type "**Computing infrastructure models**".

Threat: Generation and use of rogue certificates

Device signing and media encryption can be critically undermined by the use of rogue certificates allowing attackers the access to Big Data assets and communication links[77]. These can then be used to access data storage and thus causing data leakage, intercept and hijack individuals' secure Web-based communications, misuse of brand, and upload/download malware or force updates, which potentially contain undesired functionality for Big Data software and hardware components.

Social networks such as Facebook are affected. According to reports in some circumstances download flaws allowed attackers to plant a malicious file on a victim's machine that looks like it is coming from a trusted Facebook domain[78].

Many assets are targeted by this threat: including asset groups "**Data**" and "**Big Data analytics**", and assets "**Software**" and "**Hardware**".

Threat: Misuse of audit tools / Abuse of authorizations / Unauthorized activities

Audit information is necessary to ensure the security of the system and understand what went wrong; it is also necessary due to compliance and regulation. The scope and the granularity of the audit might be different in a Big Data context and the effect of the misuse of such information may be amplified.

For example, key personnel at financial institutions require access to large data sets that contain personally identifiable data . Also, there can be massive breaches of privacy when employees of providers hosting social networks, using their administrative credentials, regularly access private user information[79]. For this reason, it is important to keep security-relevant chronological records. Since the misuse and abuse

---

[75] See "How to Refine and Visualize Server Log Data" by is described by computer software company Hortonworks, http://hortonworks.com/hadoop-tutorial/how-to-refine-and-visualize-server-log-data/, accessed December 2015.
[76] Cloud Security Alliance (CSA) Big Data Working Group, Expanded Top Ten Big Data Security and Privacy Challenges, April 2013. See https://cloudsecurityalliance.org/research/big-data/ (last visited: September 2015).
[77] Christopher Soghoian, Sid Stamm, "Certified Lies: Detecting and Defeating Government Interception Attacks against SSLSSL" (Short Paper), Center for Applied Cybersecurity Research, Indiana University, 2011.
[78] See Kaspersky Lab blog, https://threatpost.com/facebook-users-open-to-attack-via-several-security-bugs/111572/
[79] See "Google fires employees for breaching user privacy" in TechSpot news, (Sept 2010) in http://www.techspot.com/news/40280-google-fired-employees-for-breaching-user-privacy.html, accessed December 2015.

of authorization can become a common issue, it is necessary to protect a large number of assets containing granular audits, documentation of the security policies, logs and cryptographic keys (e.g. all the assets included in category "Security and privacy techniques" of our asset taxonomy).

The assets targeted by these threats include "**identification record data**", "**Database management systems (DBS)"** (for example "**NoSQL**" and "**NewSQL**") and asset group "**Security and privacy techniques**".

Threat: Failures of business process

Failures of business process according to ENISA taxonomy are threats of damage and/or loss of assets due to improperly executed business process. In Big Data, this class includes all threats related to data integrity that can be favoured by Big Data storage policies. In particular, the highly-replicated and eventual consistency nature of big data represents a driver towards attacks to data integrity, where data items stored in different replicas can be inconsistent. This scenario is summarized in the new concept of ***Big Data degradation***, which represents an increasing risk for Big Data correctness.

This scenario also defines a "Big Data Leak", a total or partial disclosure of a Big Data asset at a certain stage of its lifecycle as opposed to a "Big Data breach" (e.g. a theft of an asset executed by breaking into the infrastructure). In our case Big Data can be unwillingly disclosed by the owner to the provider of an outsourced process, for example when computing data analytics[51]. This disclosure of information, at a certain stage of the Big Data lifecycle, can be exploited by an honest, but curious attacker, even without hostile intention.

Also, several cases of inadequate anonymisation of users are reported. While data collection and aggregation uses anonymization techniques, individual users can be re-identified by leveraging other Big Data datasets, often available in the public domain[80]. This is an emergent phenomenon introduced by Big Data variety that has the ability to infer identity from anonymized datasets by correlating with apparently innocuous public information. Examples related to de-identification of personally identifiable information (PII) are given by the AOL case[81] and by NIST Big Data publications in Web logs collection and analysis[82]. For a more detailed study on deanonymizatuion and anonymity issues in Big Data systems see ENISA's report "Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics ".

The assets targeted by this threat include asset groups "**Data**" and "**Big Data analytics**".

4.2.4 Threat Group: Legal

This group includes threats due to the legal implications of a Big Data system such as violation of laws or regulations, the breach of legislation, the failure to meet contractual requirements, the unauthorized use of Intellectual Property resources, the abuse of personal data, the necessity to obey judiciary decisions and court orders.

---

[80] Sabrina De Capitani di Vimercati, Sara Foresti, Giovanni Livraga, Pierangela Samarati, "Data Privacy: Definitions and Techniques," in International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 20, n. 6, pp. 793-817 (December 2012).

[81] See https://en.wikipedia.org/wiki/AOL_search_data_leak, accessed December 2015.

[82] See NIST Big Data Interoperability Framework: Volume 4, Security and Privacy. Use case: Web traffic analytics in retail and marketing.

Threat: Violation of laws or regulations / Breach of legislation / Abuse of personal data

Data storage in the European Union falls under the Data Protection directive: organizations are required to *i)* adhere to this compliancy law throughout the life of the data, *ii)* remain responsible for the personal data of their customers and employees, and *iii)* guarantee its security even when a third-party like a cloud provider processes the data on their behalf.

In the traditional data centric model, data is stored on-premise, and every organization has control over the information. In Big Data, instead, a real concern is arising about the security of this massive amount of digital information and the protection of the critical infrastructure supporting it, as demonstrated by a vast literature about privacy risks[83] [84] [85] [86].

We should also note that EU has stricter regulations regarding the collection of personal data than other countries, but sometimes multinationals operating in the EU are based in the United States. In this context, the most important privacy issues are how to protect individual privacy when the data is stored in multiple sites, and how efficient the protection is[Error! Bookmark not defined.].

Big Data also raises the potential issue of data residency[87]. Data, when stored in cloud storage of providers that offer multi-national storage solutions, may fall under different legal jurisdictions. An example brought by the NIST Big Data Public Working Group regards the custody of pharmaceutical data beyond trial disposition, which is unclear, especially after firms merge or dissolve[88].

The assets targeted by this threat include asset groups "**Data**" (especially "**identification record data**") and "**Roles**".

---

[83] Venkat N. Gudivada (Marshall University), Ricardo Baeza-Yates (Yahoo Labs), Vijay V. Raghavan (University of Louisiana), "Big Data: Promises and Problems", Issue No.03, 2015, Published by the IEEE Computer Society. See http://www.computer.org/csdl/mags/co/2015/03/mco2015030020.html. The book asserts that "veracity—due to intermediary processing, diversity among data sources and in data evolution raises concerns about security, privacy, trust, and accountability, creating a need to verify secure data provenance".

[84] White House Big Data Report, published on May 1, 2014. See https://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf, accessed December 2015.

[85] See various examples made by Raymond Chi-Wing Wong, "Big Data Privacy", in Journal of Information Technology & Software Engineering, Department of Computer Science and Engineering, Hong Kong University of Science and Technology, China, http://www.omicsgroup.org/journals/big-data-privacy-2165-7866.1000e114.php?aid=10289. The article cites different use cases: *(i)* individuals of a medical dataset were identified due to the insufficient privacy protection, *(ii)* datasets including search logs were released by an internet American provider but it was possible to identify a person using several individual-specific queries, *(iii)* a popular online movie rental service with a recommender system, proposed movies to its customers based on their predicted movie preferences, however almost all of the subscribers could be uniquely identified, *(iv)* many mobile clients using location-based services (LBS) had serious privacy concerns about disclosing their locations together with their personal information.

86 S. De Capitani di Vimercati, S. Foresti, P. Samarati, "Managing and Accessing Data in the Cloud: Privacy Risks and Approaches," in Proc. of the 7th International Conference on Risks and Security of Internet and Systems (CRiSIS 2012), Cork, Ireland.

[87] See Storing Data In The Cloud Raises Compliance Challenges in http://www.forbes.com/sites/ciocentral/2012/01/19/storing-data-in-the-cloud-raises-compliance-challenges/, accessed December 2015.

[88] See NIST Special Publication 1500-4. Use case: Pharmaceutical clinical trial data sharing.

4.2.5 Threat Group: Organisational threats

This group includes threats pertaining to the organizational sphere.

Threat: Skill shortage

The analysis of large datasets can underpin new waves of productivity growth and innovation, and unlock significant value. However, companies and policy makers must tackle significant hurdles, such for instance a possible shortage of skilled data scientists and managers[89].

The asset targeted by this threat is asset group "**Roles**".

[89] See for example reports from McKinsey http://www.mckinsey.com/features/big_data and from the Financial Times http://www.ft.com/cms/s/0/953ff95a-6045-11e4-88d1-00144feabdc0.html#axzz3ntU3lM00, accessed December 2015.

5. Threats agents

According to ENISA Threat Landscape 2013[90], a threat agent is "*someone or something with decent capabilities, a clear intention to manifest a threat and a record of past activities in this regard*". For Big Data asset owners it is crucial to be aware of which threats emerge from which threat agent group. This study does not develop a new glossary on threat agents, but utilises the ENISA Threat Landscape 2013's consolidation of several publications[91].

The categorization of threat agents is as follows:

Corporations: they refer to organizations/enterprises that adopt and/or are engaged in offensive tactics. In this context, corporations are considered as hostile threat agents and their motivation is to build competitive advantage over competitors, who also make up their main target. Depending on their size and sector, corporations usually possess significant capabilities, ranging from technology up to human engineering intelligence, especially in their area of expertise.

Cyber criminals: they are hostile by nature. Moreover, their motivation is usually financial gain and their skill level is, nowadays, quite high. Cybercriminals can be organised on a local, national or even international level.

Cyber terrorists: they have expanded their activities and engage also in cyber-attacks. Their motivation can be political or religious, and their capability varies from low to high. Preferred targets of cyber terrorists are mostly critical infrastructures (e.g. public health, energy production, telecommunication), as their failures cause severe impact in society and government. It has to be noted, that in the public material analyses, the profile of cyber terrorists still seems to be blurred.

Script kiddies: they are unskilled individuals using scripts or programs developed by others to attack computer systems and networks, and deface websites.

Online social hackers (hacktivists): they are politically and socially motivated individuals that use computer systems to protest and promote their cause. Their typical targets are high profile websites, corporations, intelligence agencies and military institutions.

Employees: they refer to the staff, contractors, operational staff or security guards of a company. They can have insider access to company's resources, and are considered as both non-hostile threat agents (i.e. distracted employees) and hostile agents (i.e., disgruntled employees). This kind of threat agents possesses a significant amount of knowledge that allows them to place effective attacks against assets of their organization.

[90] "ENISA Threat Landscape 2013", see https://www.enisa.europa.eu/activities/risk-management/evolving-threat-environment/enisa-threat-landscape/enisa-threat-landscape-2013-overview-of-current-and-emerging-cyber-threats, accessed December 2015.

[91] For example the Cyber Security Assesment Netherlands (see https://www.ncsc.nl/english/current-topics/news/cyber-security-assesment-netherlands.html), the Verizon report (see http://www.verizonenterprise.com/DBIR/)

Nation states: they can have offensive cyber capabilities and use them against an adversary. Nation states have recently become a prominent threat agent due to the deployment of sophisticated attacks that are considered as cyber weapons. From the sophistication of these malware, it can be confirmed that Nation states have a plethora of resources and they have a high level of skills and expertise[92].

All agents listed in this section, may have an interest in exploiting certain vulnerabilities in Big Data for different reasons. Only some specific threats come more typically from certain agents, as, for instance, the abuse of authorization that is related to corporation employee, who can use their administrative credentials to access systems. In the following table we propose a cross relation between threats and agents in Big Data. Annex C presents an overall mappings between assets, threat agents and threats.

[92] Nation-state-sponsored attacks happen more often than it is believed. For example Kaspersky Lab revealed that it had discovered an infiltration in several of its internal systems and that the attack was believed to be sponsored by a Nation state. See http://www.crn.com/slide-shows/security/300077563/the-10-biggest-data-breaches-of-2015-so-far.htm/pgno/0/2, accessed December 2015.

	CORPORATIONS	CYBER CRIMINALS	CYBER TERRORISTS	SCRIPT KIDDIES	ONLINE SOCIAL HACKERS	EMPLOYEES	NATION STATES
Unintentional damage / loss of information or IT assets							
Information leakage/sharing due to human error						●	
Leaks of data via Web applications (unsecure APIs)	●	●	●		●		●
Inadequate design and planning or improperly adaptation						●	
Eavesdropping/ Interception/ Hijacking							
Interception of information	●	●	○		●	○	●
Nefarious activity/Abuse							
Identity fraud	●	●	●	●	●	●	●
Denial of service	●	●	●				●
Malicious code / software / activity	●	●	●			●	●
Generation and use of rogue certificates	●	●	●				●
Misuse of audit tools / Abuse of authorizations / Unauthorized activities						●	
Failures of business process						●	
Legal							
Violation of laws or regulations / Breach of legislation / Abuse of personal data	○					●	●
Organisational							
Skill shortage						◌	

●: Denotes main threat agents exploiting said threat

○: Denotes potential secondary agents exploiting said threat

◌: Denotes agents is affected by said threat

Table 1: Involvement of threat agents in threats

6. Good practices

In this section, we provide a discussion summarizing good practices[93] to protect Big Data assets. A good practice is a method or technique that has consistently shown results superior to those achieved with other means, and that is used as a benchmark. To this aim, different sources have been collected, reviewed, and mapped to the previously identified Big Data threats. They specify vulnerabilities, recommendations, controls, countermeasures, and good practices published by institutions or working groups, and relevant for the protecting the assets and counteracting the threats in this report. The first result of our analysis is that publicly available information on Big Data security issues mainly originates from research and is based on requirements and generic assumptions, while materials of real-life experience are not often available. This is mainly due to the fact that development of Big Data infrastructures and their related security measures are at an early stage of maturity. In fact, on one side, many of Big Data infrastructures have been operational for a limited period of time; on the other side, Big Data security assessment is in many cases managed confidentially for reasons of competitiveness.

Generally speaking, Big Data being a collection of input channels from sensors, networks, storage and computing systems, and output to data consumers, there is shared responsibility for security and infrastructure management. Every party, such as a data provider or a data consumer, should be conscious that its own security also depends on the security of its neighbours. Countermeasures and good practices are expected to be implemented to increase security of single parties, and of other related parties when applicable.

Different documents produced by the following bodies have been examined: ISO[94], COBIT[95], Council on Cyber Security (CCS)[96] and NIST[97]. ISO terminology proposes security controls, while COBIT provides best practices that allow bridging the gap between control requirements, technical issues and business risks. The CCS is an independent and not-for-profit organization, which presents a recommended set of actions (the so called CIS Critical Security Controls for Effective Cyber Defence). When appropriate, we provide practices suggested by the NIST Big Data use cases. During the analysis, we tried to uniform the terminologies used by the above bodies, which in some cases were nonhomogeneous. For controls and technologies specifically directed towards data protection see ENISA's "Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics" (2015).

One more source of potential controls and technical countermeasures stems from the use of Big Data analytics as a tool for increasing system and data security, and improving intrusion detection and prevention. For completeness a small presentation of the expected capabilities is given in Annex E: Big Data analytics for security.

---

[93] This kind of terminology varies from body to body, for example ISACA, ISF and ANSSI propose good practices (or "bonnes pratiques"), ISO suggests security controls, NIST proposes safeguards/countermeasures, while the (German) Federal Office for Information Security (BSI) suggests safeguards.
[94] International Organization for Standardization. See http://www.iso.org/iso/home.html, accessed December 2015.
[95] Control Objectives for Information and Related Technology (COBIT). See http://www.isaca.org/cobit/pages/default.aspx, accessed December 2015.
[96] Council on Cyber Security (CCS). See http://www.counciloncybersecurity.org/about-us/, accessed December 2015.
[97] National Institute of Standards and Technology. See http://www.nist.gov, accessed December 2015.

THREAT GROUP	THREAT	ASSET GROUP / ASSET TYPE /ASSET AFFECCTED BY THEAT	MITIGATING GOOD PRACTICES, THREATS MITIGATION EXAMPLES AND CONSIDERATIONS ABOUT BIG DATA	ASSETS NOT OR PARTIALLY COVERED, AND OTHER GAPS
Unintentional damage / loss of information or IT assets	Information leakage/sharing due to human error	Data, Applications and Back-end services	ISO 27001 suggests the use of cryptography[98] to deal with unintentional leakages and prevent unauthorized access to sensitive data and systems. However, encryption key management can be challenging in Big Data[99]. According to NIST Big Data publications, security for cryptographic keys takes on "additional complexity". This is due to more consumer-provider relationships and greater demands and variety of infrastructures "on which both the key management system and protected resources are located". Issues related to the use of cryptography in Big Data are: *i)* how to protect sensitive information maintaining performance, *ii)* to allow the protection of not only files and disks, but also of logical and physical fragments. For example protection can be achieved through the encryption of data blocks, which works particularly well when Hadoop is running. Some ad hoc solutions, such as data encryption and key management tools to secure Big Data vaults, are provided by the industry[100], iii) some Big Data frameworks can't support encryption without compromising their inherent scalability and performance[101]. As a good practice, NIST Big Data Working Group publications suggest that "encryption keys should be managed by chief security officers (CSO) only and that separate key pairs should be issued for customers and internal users"[102].	Applications and Back-end services only partially covered: cryptography requires specific tools and there are issues of scalability and performance Other gaps: Roles (administrative roles become critical)

[98] "Policy on the use of cryptographic controls" and "key management" are countermeasures to avoid information leakage/sharing. These countermeasures have the objective to "ensure proper and effective use of cryptography to protect the confidentiality, authenticity and/or integrity of information" (ISO 27001).

[99] ENISA "Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics" (2015)

[100] Specialized companies provide ad hoc Big Data solutions, for example Cloudera Gazzang in the field of key management, see https://gigaom.com/2014/06/03/cloudera-acquires-big-data-encryption-specialist-gazzang/, accessed December 2015.

[101] "Trustworthy Processing of Healthcare Big Data in Hybrid Clouds" in "Computing edge: Big Data", IEEE Xplore Digital Library, 2015. See http://www.computer.org/cms/Computer.org/computing-edge/ce-nov15-final.pdf, accessed December 2015.

[102] NIST Big Data use case: "NIELSEN HOMESCAN: PROJECT APOLLO" in NIST Special Publication 1500-4

THREAT GROUP	THREAT	ASSET GROUP / ASSET TYPE /ASSET AFFECCTED BY THEAT	MITIGATING GOOD PRACTICES, THREATS MITIGATION EXAMPLES AND CONSIDERATIONS ABOUT BIG DATA	ASSETS NOT OR PARTIALLY COVERED, AND OTHER GAPS
			However, there could be other gaps: for example some roles, such as security officers and system administrators, become critical in such contexts due to the fact that they could access systems with full privileges.	
	Leaks of data via Web applications (unsecure APIs)	Data, Storage Infrastructure models	Big Data leakages via Web applications, such as unsecure APIs, or inadequate design/planning, or improper software adaptation, can be handled only with better security design, because development is at an early stage. Cryptography can be used for data protection, but there are some limitations (see above). ISO 27001 proposes good practices as "security in development and support processes" with the objective "to ensure that information security is designed and implemented within the development lifecycle of information systems". NIST Big Data publications propose "regular data integrity checks[103] to avoid injections" to APIs, as a good practice for the recently developed non-relational data stores.	Computing Infrastructure models and Storage Infrastructure models only partially covered: tool design to be improved
	Inadequate design and planning or improperly adaptation	Data, Software, Computing Infrastructure models, Storage Infrastructure models, Big Data analytics		
Eavesdropping, Interception and Hijacking	Interception of information	Data, Applications and Back-end services, Roles	Countermeasures proposed by ISO 27001 are cryptography ("policy on the use of cryptographic controls" and "key management") and "network security management" (with the objective "to ensure the protection of information in networks and its supporting information processing facilities"). However, as discussed above in this document, encryption key management might be difficult to handle in Big Data. The use of ad hoc key management tools is advisable. In some cases, protection of Big Data via centralized cryptography systems could be difficult to achieve. For example, when there are data streams that originate from a very large number of sensors. Centralized	Streaming data: some additional protection is necessary when centralized systems are not applicable Applications and Back-end services only partially covered: need to use specific protective

[103] NIST Big Data use case: "NIELSEN HOMESCAN: PROJECT APOLLO" in NIST Special Publication 1500-4

THREAT GROUP	THREAT	ASSET GROUP / ASSET TYPE /ASSET AFFECCTED BY THEAT	MITIGATING GOOD PRACTICES, THREATS MITIGATION EXAMPLES AND CONSIDERATIONS ABOUT BIG DATA	ASSETS NOT OR PARTIALLY COVERED, AND OTHER GAPS
			systems for a large number of entities is very challenging given the real-time requirements and the effect on the network performance[104]. A good practice is to extend methodologies such as the Trusted Platform Model (TPM). The enforcement through the use of trusted platforms, such as TPM, is also suggested by NIST Big Data publications[105].	tools for some critical services
Nefarious Activity/Abuse	Identity fraud	Identification record data, Applications and Back end services (such as, for example, billing services), Servers	According to NIST Big Data publications "access control is one of the most important areas of Big Data" and "one overarching rule is that the highest classification of any data element or string governs the protection of the data"[106]. ISO 27001 proposes "information classification" with the objective "to ensure that information receives an appropriate level of protection in accordance with its importance to the organisation". This is a general good practice that helps identifying the data to be protected. If data is accessed from, or transmitted to the cloud, Internet, or another external entity, then the data should be protected based on its classification. The use of trustworthy processing platforms is recommended; for example, Big Data on-premise infrastructure (private clouds) for data storage, and the image archiving and communication systems, if and when possible[107]. However, when using private clouds other gaps might	Identification record data and Back end services are only partially covered, since secure data protection in any circumstances is difficult to achieve

[104] See http://www.cisco.com/web/about/ac123/ac147/archived_issues/ipj_15-3/153_internet.html, accessed December 2015.
[105] NIST Big Data Interoperability Framework, Volume 4, Security and Privacy.
[106] NIST Big Data Interoperability Framework, Volume 4, Security and Privacy. Appendix B: Internal Security Considerations within Cloud Ecosystems
[107] "Trustworthy Processing of Healthcare Big Data in Hybrid Clouds" in "Computing edge: Big Data", IEEE Xplore Digital Library, 2015. See http://www.computer.org/cms/Computer.org/computing-edge/ce-nov15-final.pdf, accessed December 2015.

THREAT GROUP	THREAT	ASSET GROUP / ASSET TYPE /ASSET AFFECCTED BY THEAT	MITIGATING GOOD PRACTICES, THREATS MITIGATION EXAMPLES AND CONSIDERATIONS ABOUT BIG DATA	ASSETS NOT OR PARTIALLY COVERED, AND OTHER GAPS
			emerge, such as: limitation in scalability, higher operational costs for analytics models and software frameworks[108], data sharing[109]. In Big Data systems that incorporate payment platforms, providers must assure the protection of personal data. All the merchants who accept credit cards are requested to be compliant with strict international standards, such as the Payment Card Industry (PCI) Data Security Standard (known as PCI DSS). However, several cases of identity frauds due to traffic capture have been recorded in recent years[110].	
	Denial of service	Servers (Virtualized Data Centres, Physical Machines, Virtual Machines), Network	The ISP or cloud provider, which host Big Data, should implement prevention controls and the user organization's security professionals should insist that ISPs take steps to install (D)DoS prevention measures. Security countermeasures at this level, for example, might include ingress filtering, rate limiting, reverse address lookup and network traffic monitoring, general DNS good practices[111], and so forth. Also,	

[108] From "Trustworthy Processing of Healthcare Big Data in Hybrid Clouds": Analytics models and software frameworks required to manage heterogeneous data might not be available in the private cloud because of higher operational costs. In general public clouds support the most commonly used analytics models and software frameworks because of their commercial interests, while private clouds deploy tools developed in-house.

[109] From "Trustworthy Processing of Healthcare Big Data in Hybrid Clouds": Another limitation is data sharing. Data must be shared with collaborators who don't have access to private clouds or who reside outside the perimeter defences. For example, a medical practitioner from a hospital in a different jurisdiction might not be able to access the data stored in the private cloud because at present, healthcare providers are generally subject to exacting regulatory requirements to ensure the security and privacy of patient and other sensitive data.

[110] See for example "2015 Data Breach Investigations Report", released by Verizon in http://www.verizonenterprise.com/DBIR/2015/ , "2015 Identity Fraud Study", released by Javelin Strategy & Research in https://www.javelinstrategy.com/news/1556/92/16-Billion-Stolen-from-12-7-Million-Identity-Fraud-Victims-in-2014-According-to-Javelin-Strategy-Research/ , Bloomberg blog in http://www.bloomberg.com/bw/articles/2014-03-13/target-missed-alarms-in-epic-hack-of-credit-card-data , description of the largest identity theft event ever recorded (TJX Companies), in http://www.computerworld.com/article/2544306/security0/tjx-data-breach--at-45-6m-card-numbers--it-s-the-biggest-ever.html, accessed December 2015.

[111] Such as Response Rate Limiting for operators of authoritative name server, disabling open recursion on name servers, accepting only DNS queries from trusted sources, etc.

THREAT GROUP	THREAT	ASSET GROUP / ASSET TYPE /ASSET AFFECCTED BY THEAT	MITIGATING GOOD PRACTICES, THREATS MITIGATION EXAMPLES AND CONSIDERATIONS ABOUT BIG DATA	ASSETS NOT OR PARTIALLY COVERED, AND OTHER GAPS
			manufacturers and configurators of network equipment should take steps to secure all devices, keeping them up-to-date. NIST Big Data Working Group provides countermeasures to deal with (D)DoS attacks in some specific use cases: • On premises (private) cloud infrastructure when applicable[112] (for example in the pharmaceutical industry). • Anti-jamming e-measures[113] in military applications. • Mitigation through combinations of traffic analytics and correlation analysis[114] As a general rule, Big Data environments should rely on the security of their ISP or cloud provider, when a public infrastructure is used. Big Data analytics could help protecting Big Data. Big Data benchmarks could help in identifying assets to be first protected by (D)DoS attacks[115] [116] [117].	

[112] NIST Big Data use case:: Pharma Clinic Trial Data Sharing.
[113] NIST Big Data use case: "MILITARY: UNMANNED VEHICLE SENSOR DATA" in NIST Special Publication 1500-4. The anti-jamming e-measures are used by US Dept. of Defence.
[114] NIST Big Data use case: network protection
[115] Big Data benchmarks could help in identifying assets to be first protected by (D)DoS attacks.[C.A. Ardagna, E. Damiani, F. Frati, D. Rebeccani, "A Configuration-Independent Score-Based Benchmark for Distributed Databases," in IEEE Transactions on Services Computing (TSC), 2015.
[116] B. Cooper, A. Silberstein, E. Tam, R. Ramakrishnan, and R. Sears, "Benchmarking cloud serving systems with ycsb," in Proc. of the ACM Symposium on Cloud Computing (SoCC 2010), Indianapolis, IN, USA, March 2010.
[117] A. Ghazal, T. Rabl, M. Hu, F. Raab, M. Poess, A. Crolotte, and H.-A. Jacobsen, "Bigbench: Towards an industry standard benchmark for big data analytics," in Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (SIGMOD 2013), New York, NY, USA, June 2013.

THREAT GROUP	THREAT	ASSET GROUP / ASSET TYPE /ASSET AFFECCTED BY THEAT	MITIGATING GOOD PRACTICES, THREATS MITIGATION EXAMPLES AND CONSIDERATIONS ABOUT BIG DATA	ASSETS NOT OR PARTIALLY COVERED, AND OTHER GAPS
	Malicious code / software / activity	Computing Infrastructure Model, Storage Infrastructure models	For protection from malware, ISO 27001[118] and COBIT[119] propose appropriate technical vulnerability management and user awareness. This applies to Big Data as well: first of all, manufacturers and configurators of equipment should take steps to secure all devices, for example keeping them up-to-date by patching flaws. However, patch management in a Big Data heterogeneous environment might be difficult. Also user awareness through education and training is advisable.	Software tools and models are only partially covered Other gaps: Roles (administrators and users must be involved)
	Generation and use of rogue certificates	Data, Software, Hardware, Big Data analytics	This can be a common issue in a cloud infrastructure. A general good practice suggests the use of strong hashing functions such as SHA-256 or SHA-512, instead of weaker and collision prone MD5 hashing algorithm[120]. Another countermeasure that could be used to prevent attacks from potentially rogue certificate authorities involves only enabling (or whitelisting) the necessary certificate authorities that are deployed in Web browsers used by your organization. However, this may require significant effort both in terms of discovering what those certificates are and disabling all the others[121].	

[118] ISO 27001 proposes different levels of security controls such as "detection, prevention and recovery controls to protect against malware shall be implemented, combined with appropriate user awareness", "technical vulnerability management" with the objective "to prevent exploitation of technical vulnerabilities", and "rules governing the installation of software by users shall be established and implemented".
[119] COBIT Align, Plan and Organise, Audit/Assurance Program: best practices are "detection, prevention and recovery controls to protect against malware shall be implemented, combined with appropriate user awareness".
[120] See https://capec.mitre.org/data/definitions/459.html, accessed December 2015.
[121] See "SSL vulnerabilities: Trusted SSL certificate generation for enterprises" in http://searchsecurity.techtarget.com/tip/SSL-vulnerabilities-Trusted-SSL-certificate-generation-for-enterprises, accessed December 2015.

THREAT GROUP	THREAT	ASSET GROUP / ASSET TYPE /ASSET AFFECCTED BY THEAT	MITIGATING GOOD PRACTICES, THREATS MITIGATION EXAMPLES AND CONSIDERATIONS ABOUT BIG DATA	ASSETS NOT OR PARTIALLY COVERED, AND OTHER GAPS
			NIST Big Data publications provide mitigation examples based on end-point input validation[122].	
	Misuse of audit tools / Abuse of authorizations / Unauthorized activities	Identification record data, Database management systems (NoSQL, NewSQ), Security and privacy techniques	Abuse of authorizations is a common security issue, often amplified in a Big Data environment. ISO 27001 good practices suggest "business requirements of access control", "user access management", and "system and application access control"[123]. However, differently from the traditional security schemes with data access/ownership built on role-based access capabilities[124], most Big Data infrastructures offer access limitations at the schema level, but no finer granularity. A good practice in Big Data can be to logically mimic row-level access control and other advanced capabilities. This often requires specific development of these functions in the applications and in the data storage management systems.	
	Failures of business process	Data (especially Identification record data), Big Data analytics	Big data brings new challenges to privacy. ISO controls propose models, methodology, and tools for providing i) privacy protection of personal information (ISO/IEC 15944) and ii) pseudonymization[125] techniques that allow for the removal of an association with a data subject [ISO/TS	

[122] NIST Big Data use case: "HEALTH INFORMATION EXCHANGE" in NIST Special Publication 1500-4: strong authentication (see X.509v3 certificates), potential leverage of SAFE (Signatures & Authentication for Everything) bridge in lieu of general PKI

[123] According to ISO 27001 "business requirements of access control" has the objective "to limit access to information", "user access management" has the objective "to ensure authorized user access and to prevent unauthorized access to systems and services", and "system and application access control" has the objective "to prevent unauthorized access to systems and applications".

[124] Most traditional security schemes have data access/ownership built on role-based access capabilities, for example relational platforms include roles, groups, schemas, label security, and various other facilities for limiting user access to authorised subsets of the available data.

[125] Pseudonymization differs from anonymization in that it allows for data to be linked to the same person across multiple data records or information systems without revealing the identity of the person. The technique is recognized as an important method for privacy protection of personal information in the health sector. It can be performed with or without the possibility of re-identifying the subject of the data (reversible or irreversible pseudonymization).

THREAT GROUP	THREAT	ASSET GROUP / ASSET TYPE /ASSET AFFECCTED BY THEAT	MITIGATING GOOD PRACTICES, THREATS MITIGATION EXAMPLES AND CONSIDERATIONS ABOUT BIG DATA	ASSETS NOT OR PARTIALLY COVERED, AND OTHER GAPS
			25237]. These methodologies, if properly implemented in Big Data, could prevent leakages for identification record data. Also "privacy by design" techniques suggest good practices such as data minimization and retention, anonymization and related de-identification methods[Error! Bookmark not defined.].	
Legal	Violation of laws or regulations / Breach of legislation / Abuse of personal data	Data (especially identification record data), Roles	Data stored in different countries may be subject to different jurisdictions. NIST Big Data Working Group publications suggest "data residency" as a requirement for cloud-based installations[126]. This suggests the disabling of automatic replications to different regions[127], storing all data in a single national location. ISO 27001 proposes compliance[128] with legal and contractual requirements in accordance with business requirements and relevant laws and regulations. COBIT[129] proposes to "determine, document, and implement physical and logical system audit and log records". We have also to note that "private clouds are inherently trustworthy"[130]. NIST proposes also awareness and training[131] of staff.	Roles: users must be conscious of legal implications
Organisational	Skill shortage	Roles	Good practices from ISO, COBIT and other organizations apply. For example ISO 27001 proposes "information security awareness, education and training", and that "appropriate contacts with special interest groups or other specialist security forums and professional associations shall be	Roles: not all the roles will be covered (for example data scientists)

[126] NIST Big Data Interoperability Framework, Volume 4, Security and Privacy. Appendix B: Internal Security Considerations within Cloud Ecosystems
[127] Cross-region replication is often enabled by default for cloud installations.
[128] "Compliance with legal and contractual requirements" has the objective "to avoid breaches of legal, statutory, regulatory or contractual obligations related to information security and of any security requirements". Management direction for information security with the objective "to provide management direction and support for information security in accordance with business requirements and relevant laws and regulations" (ISO 27001).
[129] COBIT Align, Plan and Organise. Audit/Assurance Program
[130] "Trustworthy Processing of Healthcare Big Data in Hybrid Clouds" in "Computing edge: Big Data"
[131] See https://web.nvd.nist.gov/view/800-53/Rev4/family?familyName=Awareness%20and%20Training, accessed December 2015.

THREAT GROUP	THREAT	ASSET GROUP / ASSET TYPE /ASSET AFFECCTED BY THEAT	MITIGATING GOOD PRACTICES, THREATS MITIGATION EXAMPLES AND CONSIDERATIONS ABOUT BIG DATA	ASSETS NOT OR PARTIALLY COVERED, AND OTHER GAPS
			maintained". Also COBIT[132] focuses on awareness and training "that ensures that general users / privileged users understand roles & responsibilities and act accordingly". CCS proposes security skills assessment and appropriate training to fill gaps. NIST focuses on contacts with security groups and associations. However, there will be possible shortage for some specific technical skills[133], which need to be developed in individuals through training long-term programs. Universities need to introduce curriculum on Big data to produce skilled technicians with this expertise.	

Table 2. Good Practices and considerations about Big Data

[132] COBIT Align, Plan and Organise. Audit/Assurance Program

[133] See for example McKinsey's Business Technology Office: The next frontier for Big data competition "However, companies and policy makers must tackle significant hurdles to fully capture big data's potential - including a shortage of skilled analysts and managers" in http://www.mckinsey.com/features/big_data, accessed December 2015.

7. Gap analysis

In this section, we provide a gap analysis for those cases where further research and investigations are required in the areas of Big Data threats, security, and good practice.

This analysis aims to close the gaps highlighted in the previous section and is summarised as follows. The use of cryptography may be not always sufficient and there are obvious risks associated to administrators and security professionals with equivalent privileges. This is especially true when threats related to information leakage and/or sharing due to human errors are considered. Furthermore, leaks of data via Web applications (unsecure APIs) and inadequate design/planning or improperly adaptation need an improved design of computing and storage infrastructure models, while streaming data from sensors may have issues of confidentiality that cannot be mitigated by current solutions. Personal identifiable information is at risk even when best practices are widely followed and calls for privacy-oriented defensive approaches. Malicious code and activities pose a risk to models of computing infrastructure and storage due to the difficulties of patch management in a Big Data heterogeneous environment, while violation of laws or regulations, breach of legislation and abuse of personal data may affect final users. All these breaches requires, on one side, Big Data specific countermeasures, and, on the other side, the involvement of policy makers to reflect changes in current IT environment in EU laws and legislations. Finally, a skill shortage in roles such as data scientists is foreseen.

We categorize the gaps into four groups: gaps (i) on data, (ii) on the use of cryptography (iii) on computing and storage models and (iv) on roles (e.g. administrators, data scientist, and final users).

Gaps on data protection

Major gaps are found due to threats to privacy (e.g., the identification of personal identifiable information) and to confidentiality of sensor data streams.

As already reported in this report, several cases of identity fraud due to traffic capture and data mining have been recorded in recent years. Big Data analysis has facilitated the intrusion of privacy by strengthening common techniques and further research in this field is required. Since countermeasures, discussed in the previous section, such as anonymization did not prove to be always effective against Big Data mining, new research efforts are made to devise better controls. For example, a promising topic, actively researched, is privacy-preserving data mining[134] (PPDM). The basic idea of PPDM is to modify the data in such a way so as to perform data mining algorithms effectively without compromising the security of sensitive information contained in the data.

In addition, it is foreseeable to have streams of data from sensors certified when possible. Since centralized cryptography systems are hard to implement when a large number of sensors is involved, the use of Trusted Computing (TC) appears to be a promising technology. Trusted computing relies on Trusted Platform Modules (TPMs)[135] [136] and related hardware to prove integrity of software, processes, and data. TPM chips are not expensive and could be fitted in sensors at build time. TPM-enabled devices could

[134] Lei Xu et al., "Information Security in Big Data: Privacy and Data Mining", Department of Electronic Engineering, Tsinghua University, Beijing, China.
[135] Morris, "Trusted Platform Module" In Encyclopaedia of Cryptography and Security, Springer (2011).
[136] TPM specifications can be found at http://www.trustedcomputinggroup.org/resources/tpm_main_specification, accessed December 2015.

provide reliable data streams. However, on the server side (e.g., Big Data cloud-based installations), the use of this new technology is more challenging since hardware TPM should be adapted to virtualized environments. A researched approach is based on the notion of virtual Trusted Platform Module (vTPM)[137], which provides secure storage and cryptographic functions of TPM to applications and operating systems running in virtual machines.

Other hardware-based security technologies include the development of new processors for the embedded smart sensors[138]. These new processors include protected areas for storage of user authentication keys, as well as areas of the processor that are off-limits to unauthorized users.

Besides the above technically-oriented aspects of data protection gaps, in 2015 ENISA has conducted a privacy-oriented assessment of Big Data "*Privacy by design in big data*"[4]. In this work, more thorough privacy gaps have been identified and recommendations have been made. Highlights include: application of privacy by design, preservation of privacy by data analytics and the need for coherent and efficient privacy policies for big data. It is recommended to refer to this document in order to obtain full perspective of security and privacy issues of Big Data.

Use of cryptography in applications and back-end services

The use of cryptography in Big Data as a mitigation countermeasure can be challenging. Gaps related to the use of cryptography are mainly related to: *i)* performance and scalability, *ii)* protection of logical and physical fragments, such as data blocks.

In fact, in Big Data, cryptography adds complexity and negatively affects performance. New dedicated products and ad hoc solutions are under development, as for example the already discussed TC and TPM technologies, while some interesting new approaches to cryptography for Big Data applications as the notion of "cryptography-as-a-service" in cloud environments[139] are emerging. In recent years, there has been a lot of discussion around novel, but still rather esoteric crypto-algorithms. Homomorphic encryption[140], honey encryption[141] and other proposals could, at least in theory, provide end-to-end data protection and confidentiality. As an example, assuming the existence of a fully homomorphic crypto-scheme, one could use public Big Data systems to perform analytics – with the expected speed or accuracy losses – without ever revealing the data to anyone else, not even the computation and storage service provider. Research is still ongoing[142] but the interested reader can find a concise study of the current state of the art in ENISA's "Privacy by design in big data: An overview of privacy enhancing technologies in the era of big data analytics" **Error! Bookmark not defined.**.

---

[137] Berger et al., "vTPM: Virtualizing the Trusted Platform Module" in Proc. of USENIX-SS 2006. Vancouver, Canada, 2006.
[138] See http://www.arm.com/markets/embedded/index.php, accessed December 2015.
[139] See 11 th International Conference on Applied Cryptography and Network Security (ACNS 2013) in https://www.trust.informatik.tu-darmstadt.de/publications/publication-details/?no_cache=1&tx_bibtex_pi1[pub_id]=TUD-CS-2013-0089, accessed December 2015.
[140] A form of encryption that allows computations to be performed directly on the cyphertext. The results are themselves the encrypted equivalent of the computation of the original data.
[141] A form of encryption that generates cypher-texts that when decrypted with the wrong key material generate bogus yet plausible-looking results.
[142] See for example https://www.nsf.gov/discoveries/disc_summ.jsp?cntn_id=136673, accessed December 2015.

Gaps on computing and storage models

Computing Infrastructure and storage models in Big Data face new challenges such as the lack of standardization and portability of security controls among different open source projects (e.g., different Hadoop versions) and Big Data vendors[143], and the poor design of security features.

Often, standards do not exist or are still under development. An example of lack of standards is brought by NIST Big Data Working Group for the shipping industry, which uses Big Data in the identification, transport, and handling of items in the supply chain[144]. However, at the moment, the status of the shipped items (e.g., unique identification number, GPS coordinates, sensors information, etc.) is not passed through the entire chain. A unique identification schema is under development within an ISO technical committee.

From a security perspective, we note that in a traditional management system as, for example, in an SQL relational database, security has slowly evolved and many new controls have been proposed over the years. Unlike such solutions, the security of Big Data components has not undergone the same level of rigor or evaluation due to the immaturity of Big Data research and development.

Gaps on roles (administrators, data scientist, and final users)

As stated in the previous section, many roles can be critical in Big Data, in particular system administrators, data scientist, and users.

Big Data administrators and other privileged users are a big concern since they require access to corporate data systems when working on behalf of the cloud services provider. Moreover, they could use their grants to access key stores and other sensitive information[145]. All the data scientist positions are unlikely to be filled in the near future, while users might not always be conscious of or care about the legal implications of data storage – legal implications that will vary large and wide around the world.

Awareness, education, and training are the keys to close these gaps concerning human resourses. Some new online educational web sites are offering specialised courses in Big Data, for example the Big Data University[146] sponsored by IBM, and MIT[147]. The Big Data University is run by a community, which includes many IBM staff members, contributing voluntarily to the development of courses, and to enhancing the site; also Amazon is contributing to the initiative[148]. Other courses are available at Massive Open Online

---

[143] Ajit Gaddam, 'Securing Your Big Data Environment', *Community event: Black Hat USA*, Las Vegas, August 2015. See https://www.blackhat.com/docs/us-15/materials/us-15-Gaddam-Securing-Your-Big-Data-Environment-wp.pdf , accessed December 2015.

[144] NIST Big Data Public Working Group, NIST Special Publication 1500-4: NIST Big Data Interoperability Framework, Volume 4, Security and Privacy (draft, April 2015)

[145] Vormetric report, Trends and Future Directions in Data Security, CLOUD AND BIG DATA EDITION, 2015. See http://enterprise-encryption.vormetric.com/rs/vormetric/images/Cloud-and-BigData-Edition-2015-Vormetric-Insider-Threat-Report-Final.pdf, accessed December 2015.

[146] See http://www.ibm.com/developerworks/data/library/techarticle/dm-1205bigdatauniversity/ and http://bigdatauniversity.com, accessed December 2015.

[147] See https://mitprofessionalx.mit.edu/courses/course-v1:MITProfessionalX+6.BDx+5T2015/about, accessed December 2015.

[148] Some courses in Big Data University are sponsored by Amazon Web Services, which provide a credit to learn Big Data on their cloud.

Course (MOOC) websites like Coursera[149]. But, as with ICT security[150], it will take years to fulfil industry's requirements on skilled and trained personnel.

Recommendations

The above gaps naturally result in a set of recommendations that can be classified as general recommendations, technical recommendation and recommendation on human resources.

General recommendations: they target the main Big Data stakeholders such as owner of Big Data projects and policy makers. In particular, stakeholders should depart by the assumption that a Big Data environment is simply a traditional data environment focusing on large amount of data. Big Data is more than a simple scalability problem, and management tools and risk assessment countermeasures and solutions should consider and address all 5V characterizing a Big Data environment.

This consideration is important both for policy makers specifying laws and regulations targeting current ICT environment, and stakeholders managing Big Data platforms and analytics. Especially for the latters, it becomes fundamental to evaluate *i)* the current level of security by understanding the assets covered (and not covered) by existing security measures, *ii)* the effectiveness of the application of good practices adapted from traditional security and privacy tools and techniques.

General recommendation requires a parallel standardization effort supporting the definition of proper and specific Big Data tools and legislations.

Technical recommendations: they target owners of Big Data projects and developers of corresponding products.

Following general recommendation of being Big Data specific, stakeholders should limit as much as possible the practice of adapting existing products to Big Data. Big Data introduces completely novel environments with new assets, threats, risk, and challenges. As a consequence, new products are needed to provide effective countermeasures and increase the trustworthiness of Big Data environments. Such products must be put in the Big Data life cycle after a careful evaluation, through pilots, aimed to verify and prove their correct behaviour. Success of these new products passes from a commitment by third-party vendors to apply security measures and stay focused on any updates.

Moreover, developers of Big Data products should benefit from new tools providing security and privacy functionalities by default.

To conclude, as already specified in the general recommendations, international bodies are invited to support this shift to Big Data specific security and privacy solutions by starting a gap analysis on Big Data standards, and new standardization activities according to the identified gaps.

Recommendations on human resources: they target human resources managing and using Big Data assets. As in traditional environments, in fact, human resources are one of the main sources of threats, and include users that attack a system either maliciously or accidentally. To limit these scenarios, all involved parties should focus on training of specialized professionals. Big players should support education initiatives on Big Data to raise/train tomorrow's scientists, fostering information and communication technology security awareness and training programs. Private companies and governmental bodies should

[149] See for example https://www.coursera.org/course/mmds, accessed December 2015.
[150] W. Lee & B. Rotoloni, "Emerging Cyber Threats Report 2016". Georgia Tech Information Security Center (2015)

encourage technical staff to attend offline/online courses from respected institutes to increase their competences. Final users should learn about their rights and threats to privacy attending courses and educational initiatives. Big Data administrator and other privileged users should cooperate with the international community to exchange on threats and promote the application of good practices as mitigation measures. Finally, Big Data administrator should rely on good practices, and report on their implementations choices in terms of considered assets, threat, countermeasures, and identified gaps.

Annex A: Full list of Big Data taxonomy

ASSET GROUP	ASSET TYPE	ASSET	ASSET DETAIL	EXAMPLES[151] AND COMMENTS
Data	Metadata	Schemas and Indexes		
		Stream grammar data		
	Structured data	Identification record data		Users' profiles and preferences
		Linked open data		
		Inferences and re-linking data		
		Databases		
	Semi-structured and unstructured data	Logs		System logs, transaction logs, security audit logs, web logs, network logs, test logs, etc.
		Messages and Web unformatted data		Emails, SMS, tweets, posts, Webpages, blogs, Wikis, etc.
		Files and documents		Repositories and File Systems
		Multimedia		
		Other non-textual material besides multi-media		
	Streaming data	Single medium streaming		Sensor streaming data
		Multi-media streaming		Remote sensing satellite streaming data
	Volatile data	Routing data		
		Random Access Memory (RAM)		
Infrastructure	Software	Operating Systems		
		Device Drivers		
		Firmware		
		Server Software	Web Server	

[151] Examples are provided when significant.

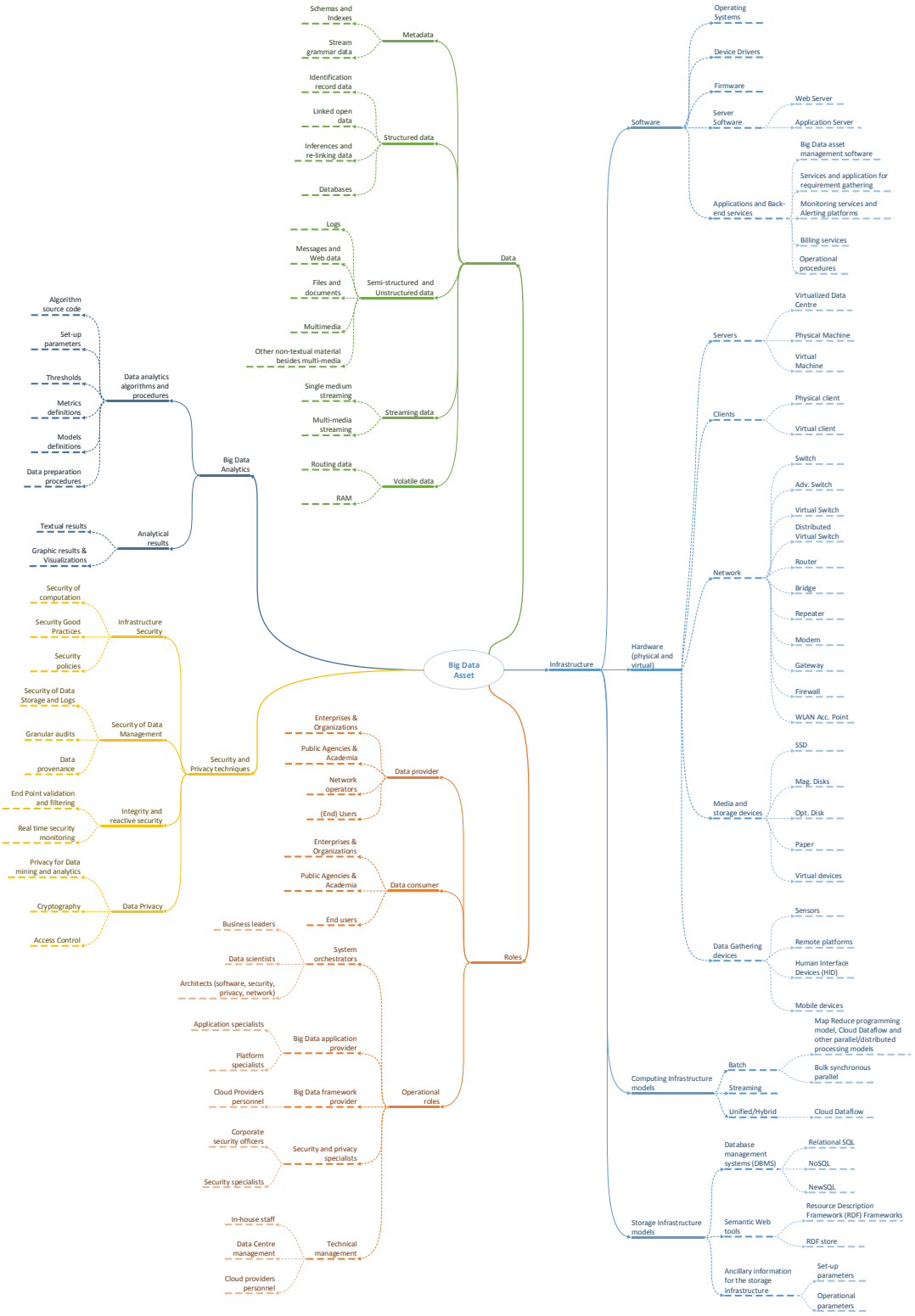
ASSET GROUP	ASSET TYPE	ASSET	ASSET DETAIL	EXAMPLES[151] AND COMMENTS
			Application Server	Database server s/w covered in Database management systems (DBMS)
		Applications and Back-end services	Big Data asset management software	Templar
			Services and application for requirement gathering	
			Monitoring services and Alerting platforms	
			Billing services	
			Operational procedures	
	Hardware (physical and virtual)	Servers	Virtualized Data Centre	VMware vSphere suite (Virtual Data Centre, Centre Management Console, etc.)
			Physical Machine	
			Virtual Machine	
		Clients	Physical client	PCs, notebooks, Mobile Devices (Tablet, Phone, PDA, etc.)
			Virtual client	
		Network	Switch	
			Adv. Switch	
			Virtual Switch	
			Distributed Virtual Switch	
			Router	
			Bridge	
			Repeater	
			Modem	
			Gateway	
			Firewall	
			WLAN Acc. Point	
		Media and storage devices	SSD	
			Mag. Disks	
			Opt. Disk	

ASSET GROUP	ASSET TYPE	ASSET	ASSET DETAIL	EXAMPLES[151] AND COMMENTS
			Paper	
			Virtual devices	
		Data Gathering devices	Sensors	
			Remote platforms	Airborne platforms, drones (UAV and RPAS), etc.
			Human Interface Devices (HID)	
			Mobile devices	
	Computing Infrastructure models	Batch	Map Reduce programming model	Apache Hadoop architecture, S4
			Bulk synchronous parallel	Apache Hama, Giraph, Pregel
		Streaming		Info sphere, Storm, Spark
	Storage Infrastructure models	Database management systems (DBS)	Relational SQL	
			NoSQL	Document oriented DBs (Mongo DB, Couch DB, Couch Base), Key Value stores In-memory (Redis, Memcache, Aerospike, etc.) and Key Value stores Dynamo-inspired (Riak, Cassandra, Voldemort, etc.), Big Table (Hbase, Cassandra), Graph-oriented (Giraph, Neo4j, Orient DB)
			NewSQL	In-memory DBS as Hstore, VoltDB
		Semantic Web tools	Resource Description Framework (RDF) Frameworks	Sesame, Virtuoso
			RDF store	Triple stores (Jena, Sesame Native, Mulgara, Virtuoso Native,), Quad RDF stores
		Ancillary information for the storage infrastructure	Set-up parameters	
			Operational parameters	
		Algorithm source code		

ASSET GROUP	ASSET TYPE	ASSET	ASSET DETAIL	EXAMPLES[151] AND COMMENTS
Big Data Analytics	Data analytics algorithms and procedures	Set-up parameters		
		Thresholds		
		Metrics definitions		
		Models definitions		
		Data preparation procedures		
	Analytical results	Textual results		
		Graphic results & Visualizations		Spatial layouts, Interactive / real time visual representations
Security and Privacy techniques	Infrastructure Security	Security of computation		
		Security Best Practices		
		Security policies		
	Security of Data Management	Security of Data Storage and Logs		
		Granular audits		
		Data provenance		
	Integrity and reactive security	End Point validation and filtering		
		Real time security monitoring		
	Data Privacy	Privacy for Data mining and analytics		
		Cryptography		
		Access Control		
5. Roles	Data provider	Enterprises & Organizations		
		Public Agencies & Academia		
		Network operators		
		(End) Users		
	Data consumer	Enterprises & Organizations		
		Public Agencies & Academia		
		End users		
	Operational roles	System orchestrators	Business leaders	
			Data scientists	

ASSET GROUP	ASSET TYPE	ASSET	ASSET DETAIL	EXAMPLES[151] AND COMMENTS
			Architects (software, security, privacy, network)	
		Big Data application provider	Application specialists	
			Platform specialists	
		Big Data framework provider	Cloud Providers personnel	
		Security and privacy specialists	Corporate security officers	
			Security specialists	
		Technical management	In-house staff	
			Data Centre management	
			Cloud providers personnel	

Annex B: Full Big Data asset taxonomy structure



Annex C: Full list of threats affecting Big Data

This section provides a list of all the threats that affect Big Data assets and that were examined in the previous sections of the document.

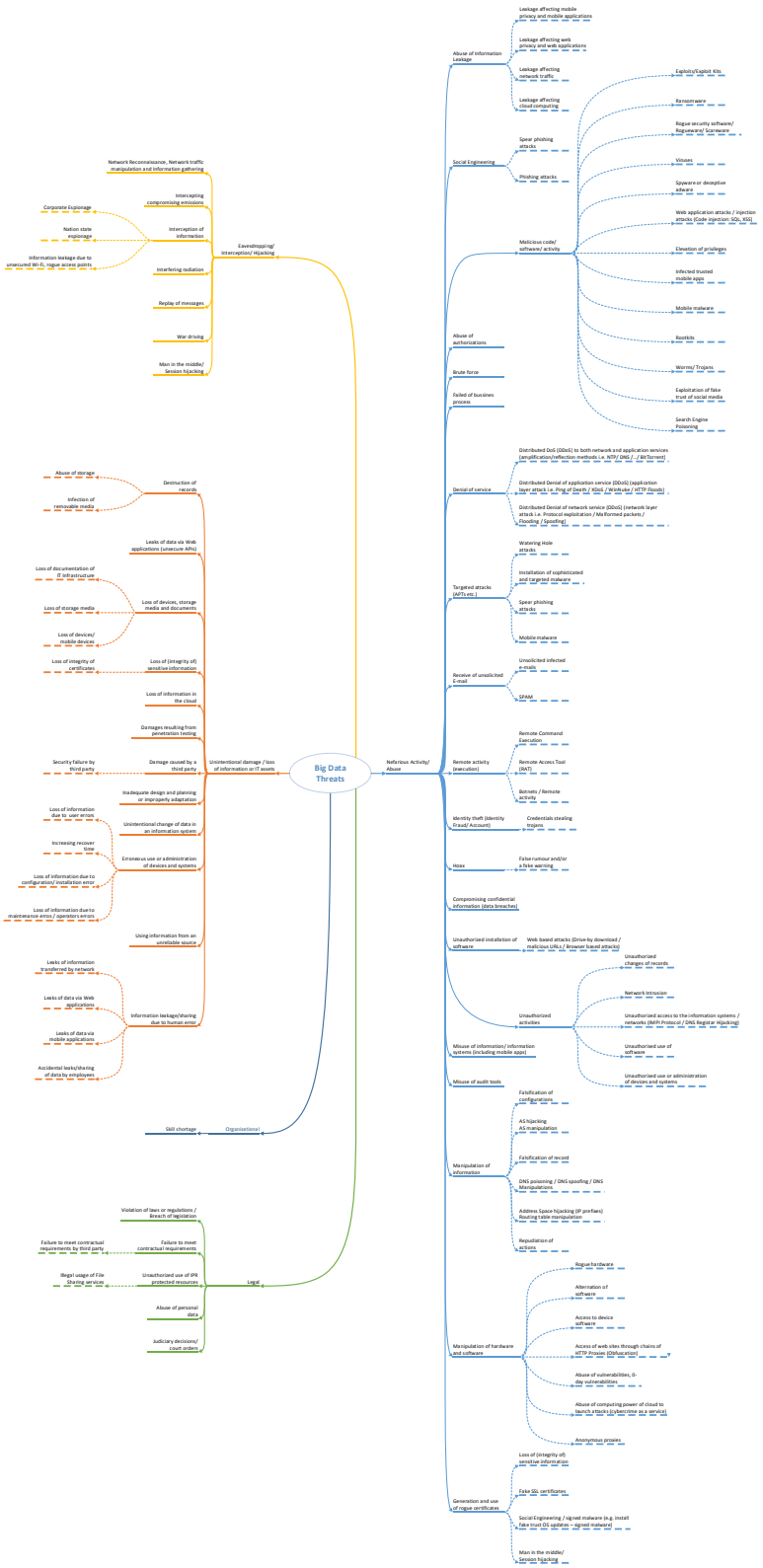
THREAT GROUP / TYPE	THREAT	THREAT AGENTS	FORSEEABLE EFFETS	AFFECTED ASSET TYPE / ASSET / ASSET DETAIL	COMMENTS
Unintentional damage / loss of information or IT assets	Information leakage/sharing due to human error	All agents applicable	Data disclosure	Data	
	Leaks of data via Web applications (unsecure APIs)	All agents applicable	Data disclosure	Data Storage Infrastructure models	
	Inadequate design and planning or improperly adaptation	All agents applicable	Data disclosure Higher probability of data leaks	Data Software Computing Infrastructure models Storage Infrastructure models Big Data analytics	Typical threat for Big Data (redundancy as a system weakness)
Eavesdropping / Interception / Hijacking	Interception of information	All agents applicable	Data disclosure	Data	
Nefarious Activity / Abuse	Identity fraud	Especially cyber criminals	Personal data disclosure	identification record data Applications and Back end services Servers	Typical threat for Big Data: the effect is amplified by the environment and can have a severe impact
	Denial of service	Especially cyber criminals and online social hackers	Distract company staff (criminals) or disable legitimate usage of websites (hacktivists)[152]	Servers	

[152] According to Global Threat Intelligence Report (GTIR) by Solutionary Inc. hacktivists utilize DDoS attacks to advance political and social objectives, disabling the legitimate usage of websites and the target's other IT resources in order to express a message of dislike or disapproval, while criminal purpose for DDoS attacks may include distracting company staff from noticing evidence of the fraudulent financial transaction, overwhelming IT with response to a serious event (allowing time for the fraudulent transaction to be completed), disabling the target organization VoIP and other IT infrastructure to disrupt communication (preventing external verification of

THREAT GROUP / TYPE	THREAT	THREAT AGENTS	FORSEEABLE EFFETS	AFFECTED ASSET TYPE / ASSET / ASSET DETAIL	COMMENTS
	Malicious code/ software/ activity	All agents applicable, but specially cyber criminals	Service disruption Data disclosure (especially financial data)	Database management systems (DBMS) Semantic Web tools Computing Infrastructure Model	Threat for new tools, such as Big Data Semantic Web tools (SPARQL, NoSQL DBs, etc.)
	Generation of rogue certificates	Especially cyber criminals	Data disclosure Service disruption	Data Software Hardware Big Data analytics	
	Misuse of audit tools / Abuse of authorizations / Unauthorized activities	All agents applicable, but especially employees for abuse of authorizations	Data disclosure	identification record data Database management systems (DBS) Security and privacy techniques"	Effects are amplified by Big Data.
	Failures of business process	N/A	Data disclosure	Data Big Data analytics	
Legal	Violation of laws or regulations / Breach of legislation / Abuse of personal data	N/A	Data disclosure	Data	Typical threat for Big Data
Organisational	Skill shortage	N/A	Low productivity growth and innovation	Roles	Typical threat for Big Data

fraudulent transfers), causing rollover of Web and application log files (destroying evidence of unauthorized intrusions)

Annex D: Full Big Data threat taxonomy structure



Annex E: Big Data analytics for security

In recent years, Big Data analytics have attracted the interests of the security community as the means to increase security protection, thanks to its promise to analyse and correlate security-related data efficiently and at unprecedented scale. Even if ICT security community has been analysing logs, network flows, system events, and other information sources to identify threats and detect malicious activities for many years, conventional technologies have not always proven to be adequate to support long-term, large-scale analytics for many reasons[153]. First, performing analytics and complex queries on large, unstructured datasets is inefficient since conventional Business Intelligence tools are designed to analyse and manage data organized in fixed and predefined schemas. Second, the management of large data warehouses is very expensive and dedicated installation, and their deployment usually requires strong business motivations.

The above limitations can be mitigated by applying Big Data analytics techniques to huge amount of backend data, to the aim of uncovering security threats, attack patterns, and security exploits. This section details how the security analytics landscape is changing with the introduction of Big Data applications and tools, such as the Hadoop framework[154], which can deploy, clean, process, analyse, and query large amounts of structured and unstructured data efficiently.

In particular, we will examine Big Data analytics as a tool for increasing system and data security, and improving intrusion detection and prevention. This falls under the more general name of threat intelligence, which is defined as "evidence-based knowledge, including context, mechanisms, indicators, implications and actionable advice, about an existing or emerging menace or hazard to assets that can be used to inform decisions regarding the subject's response to that menace or hazard" [155].

Threat intelligence evolution started years ago, security companies started building anti-virus products, later on created signatures to capture bad emails, and finally introduced reputation through behavioural indicators to infer which files were malware and which messages were spams[156]. More recently, when networks of compromised devices (botnets) could evade detection and defeat reputation, heavier reliance was put on other techniques, such as attackers' analysis with a resulting aggregation of big amounts of data, to get a better view on the threat landscape. Nowadays, threat intelligence can be considered as a knowledge base, which represents the synthesis of information detailing potential threats with a solid understanding of network structure, operations, and activities[157]. This knowledge base is a collection of indicators, often called "threat feeds", which must be contextualised with a baseline of normal network activities. Big Data security analytics can be used to improve chances of detection and ascertain trends.

---

[153] Alvaro A. Cárdenas (University of Texas, Dallas), Pratyusa K. Manadhata (HP Labs), Sreeranga P. Rajan (Fujitsu Laboratories of America), Big Data Analytics for Security, Co-published by the IEEE Computer and Reliability Societies , November/December 2013

[154] Apache Hadoop http://hadoop.apache.org, accessed December 2015.

[155] As it is defined by Gartner in https://www.gartner.com/doc/2487216/definition-threat-intelligence, accessed December 2015.

[156] Securosis, "Building an Early Warning System" (2013), see https://securosis.com/research/publication/building-an-early-warning-system, accessed December 2015.

[157] INSA Cyber Intelligence Task Force White Paper, see http://www.insaonline.org/i/d/a/Index.aspx, accessed December 2015.

Big Data analytics can adopt different approaches. For example, Hadoop can be programmed to detect all data entering and exiting the network. This configuration has been used to pick up odd activities, such as when an infected PC or server starting communicating at unusual times[158]. Another example uses Hadoop to monitor system logs[159]: a service called Flume is used for streaming log data into Hcatalog, a SQL based data store, then Pig is used to query and refine the data and Elastic Search for a high-level visualization. A proposed Early Warning System methodological framework[156] implements a systematic process to collect and aggregate security data internally, establish a number of baselines that identify normal behaviour, gather external intelligence (threat feeds) from third parties, and uses Big Data analytics to analyse this information for particular dangers.

Future directions for research in Big Data analytics for security aim at addressing business risks for the whole organisation[160]. In fact, without a risk-based approach to security, organisations could waste valuable ICT resources for vulnerabilities that are not causing in reality big threats to the business. Also it will be important to filter security Big Data to the information that is just relevant to specific stakeholders' roles and responsibilities.

In the following of this section we focus on Big Data analytical solutions for the following threats: anomaly detection, denial of service, fraud detection, and botnets.

Anomaly detection

Telecommunication providers observe Internet traffic and analyse it to study malicious activities. For this reason a large amount of data is collected and millions of global DNS requests, HTTP transactions and full packet information are correlated. Network operators analyse the data usually relying on detection tools based on anomaly detection techniques. These tools are based on a reference model of the normal traffic behaviour and compute the deviation from it. However, to cope with the global growth of Internet and obtain results in a reasonable time, the Internet traffic is usually sampled, even if sampling is inherently detrimental to anomaly detection.

Big Data tools promise to improve the analysis of these large datasets, by providing fast and accurate analytics. Recently, the research community has then put a lot of effort in the development of efficient tools using the MapReduce model. As an example, EU funded project NECOMA[161] is investigating the benefits of MapReduce to achieve real-time anomaly detection with non-sampled traffic. The goal of the project is to provide high scalable and fault tolerant tools implementing features for anomaly detection. The project proposes a MapReduce-based framework that consists of two steps[162]: first, the traffic is

[158] Company King, creator of the popular Candy Crush Saga mobile game, built its own Big Data framework to look into strange behaviour on its machines. See http://raconteur.net/technology/hacking-hackers-with-big-data, accessed December 2015.

[159] Company Hortonworks gives an example of system log monitoring in a tutorial, where DDOS attacks are being tracked down by system admins, See http://hortonworks.com/hadoop-tutorial/how-to-refine-and-visualize-server-log-data/, accessed December 2015.

[160] Alaa Hussein Al-Hamami (Amman Arab University, Jordan) and Ghossoon M. Waleed al-Saadoon (Applied Sciences University, Bahrain), Handbook of Research on Threat Detection and Countermeasures in Network Security See http://www.igi-global.com/book/handbook-research-threat-detection-countermeasures/110015, accessed December 2015.

[161] See NECOMA (Nippon-European Cyberdefense-Oriented Multilayer threat Analysis) project http://www.necoma-project.eu. NECOMA is funded by the European Union Seventh Framework Programme and by the Strategic International Collaborative R&D Promotion Project of the Ministry of Internal Affairs and Communication, Japan.

[162] Romain Fontugne, Johan Mazel, Kensuke Fukuda, "Hashdoop: A MapReduce Framework for Network Anomaly Detection" in 2014 IEEE INFOCOM Workshops: 2014 IEEE INFOCOM Workshop on Security and Privacy in Big Data

divided into splits using a hash function and preserving both the spatial and temporal structures of the traffic; second, detectors identify anomalies in each split of data.

Denial of service

A special case of anomaly detection is the discovery of Distributed Denial of Service attacks (DDoS). For example, anomaly detection tools can identify DDoS attacks by observing numerous requests sent to the same service. However, detection performance and accuracy can be highly decreased by the availability of a sampled set of requests only[161]. Also here the use of Big Data tools can help: various designs of MapReduce based frameworks[161] and of detection algorithm for the major flooding attacks[163] (TCP-SYN, HTTP GET, UDP and ICMP) are reported.

Fraud detection

Fraud is a deliberate deception to obtain unfair or unlawful gain. The purpose of a fraud may be monetary gain or other benefits, such as retrieving certificates by way of false statements. New approaches based on Big Data analytics can be used to combat fraud, by i) correlating real-time and historical account activity, ii) relying on baselines to spot abnormal user behaviour uncovering trends and patterns in large amounts of data, iii) establishing patterns and relationships, and iv) making non-obvious connections between disparate sources of data[164]. By these means, businesses can identify fraud risks at an early stage, thus preventing crime and solving investigations. The technical challenge about fraud analysis involves looking for behavioural patterns and building a profile of normal activities[165], accessing sparse financial information data and parsing unstructured text, and understanding discrepancies in customer transactions. We also notice that companies are not inclined to reveal real cases of fraud that have undergone, unless law requires them and real cases are not often reported.

The use of Big Data tools is becoming common for insolvency and forensic professional services, and for teams of dedicated investigators covering financial investigations[166]. The crimes can be traced by analysing structured and unstructured data sources such as bank statements, PDF files, emails, invoices and spread sheets. The same article claims that "technology has enabled the investigation business to move from struggling with huge amounts of information stored in spread sheets to a faster, more accurate,

---

[163] Sufian Hameed, Usman Ali, On the Efficacy of Live DDoS Detection with Hadoop, National University of Computer and Emerging Sciences (NUCES).

[164] IBM Software White Paper, Extending security intelligence with big data solutions. See http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=SA&subtype=WH&htmlfid=WGW03020USEN. The whitepaper claims that traditional security solutions are no longer sufficient to defend against new escalating threats, and proposes the company's Big Data analytical tools to deal with Internet-scale botnet discovery, full-spectrum fraud detection and comprehensive insider threat analysis.

[165] IBM Software White Paper, Extending security intelligence with big data solutions. See http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=SA&subtype=WH&htmlfid=WGW03020USEN. The whitepaper claims that traditional security solutions are no longer sufficient to defend against new escalating threats, and proposes the company's Big Data analytical tools to deal with Internet-scale botnet discovery, full-spectrum fraud detection and comprehensive insider threat analysis.

[166] "Smarter fraud investigations with big data analytics" by Shaun Hipgrave (IBM) cites the case of Griffins, a UK based insolvency and forensic services company. The company has one of the largest teams of dedicated investigators covering insolvency and financial investigations. In addition to this, they also provide services for creditors, debtors and professional advisors. They use analytics software to reduce the cost, time and complexity associated with fraud and forensic investigations for litigation, see also http://www.computerweekly.com/news/2240180084/Insolvency-firm-Griffins-speeds-up-fraud-forensics-with-IBM-analytics, accessed December 2015.

intelligence-led approach that helps to solve cases related to money laundering, missing trader fraud, and theft of company assets."

Botnets

Network flow data collected by telecommunication providers are also analysed to identify malicious communications associated with botnets. Years of historical data on DNS transactions across enterprises are analysed by Big Data tools, applying custom analytics to help in identifying suspicious domain names used by botnets[167], infected hosts, and past intrusions. Another approach uses a distributed computing framework that examines host relationships using a combination of PageRank and clustering algorithms to track the command-and-control channels in the botnet.

[167] François et al, BotCloud: Detecting botnets using MapReduce, Published in: IEEE International Workshop on Information Forensics and Security (WIFS), 2011

Annex F: Summary of threat taxonomies

In this report we have used the ENISA threat taxonomy to map threats to Big Data assets. Obviously it is not the only threat taxonomy that may be used for this exercise. In this section we provide a very short review of some other taxonomies of threats and cyber attacks presented in international conferences and workshops (**Error! Reference source not found.**) that might be used as an alternative. The review showed that there are no specific threat taxonomies for Big Data assets.

#	TAXONOMY / CATEGORISATION	MAIN CHARACTERISTICS	REFERENCE
#1	Cyber attack taxonomy (called AVOIDIT)	Identification of cyber-attacks and potential countermeasures.	Simmons, C. B., Shiva, S. G., Bedi, H., & Dasgupta, D., 'AVOIDIT: A Cyber Attack Taxonomy', Presented at the *9th ANNUAL SYMPOSIUM ON INFORMATION ASSURANCE (ASIA'14)*, JUNE 3-4, ALBANY, NY, 2014.
#2	Taxonomy of information security threats	Study on categories of security threats.	Im, G. P., & Baskerville, R. L., 'A Longitudinal Study of Information System Threat Categories: The Enduring Problem of Human Error', *SIGMIS Database*, 2005, Vol. 36, No 4, pp. 68–79. doi:10.1145/1104004.1104010
#3	Taxonomy of Distributed Denial of Services (DDoS) attack and defence mechanisms	Classification of DDoS attacks and defence strategies.	Mirkovic, J., & Reiher, P., 'A Taxonomy of DDoS Attack and DDoS Defence Mechanisms', *SIGCOMM Comput. Commun. Rev.*, 2004, Vol. 34, No 2, pp. 39–53. doi:10.1145/997150.997156
#4	Classification of network and computer attacks	Study on computer and network security as well as consistency in language with attack description.	Hansman, S., & Hunt, R., 'A Taxonomy of Network and Computer Attacks', *Comput. Secur.*, 2005, Vol. 24, No 1, pp. 31–43. doi:10.1016/j.cose.2004.06.011
#5	Taxonomy of cyber attacks and cyber adversaries	Analysis of attack taxonomies and classification cyber adversaries.	Meyers, C. A., Powers, S. S., & Faissol, D. M., *Taxonomies of Cyber Adversaries and Attacks: A Survey of Incidents and Approaches* (No. LLNL-TR-419041). Livermore, CA: Lawrence Livermore National Laboratory (LLNL), 2009.
#6	Structured or unstructured lists of threats to be coupled with attack taxonomies	Classification scheme to help developers find relevant attacks	Uzunov, A. V., & Fernandez, E. B., 'An Extensible Pattern-based Library and Taxonomy of Security Threats for Distributed Systems', *Comput. Stand. Interfaces*, 2014, Vol. 36, No 4, pp. 734–747. doi:10.1016/j.csi.2013.12.008

Article #1 proposes a cyber attack taxonomy called AVOIDIT (Attack Vector, Operational Impact, Defence, Information Impact, and Target). The study uses five major classifiers to characterize the nature of an attack: *(i)* classification by attack vector, *(ii)* classification by operational impact, *(iii)* classification by defence, *(iv)* classification by informational impact, and (v) classification by attack target. This technique is useful to associate threats to Big Data with the corresponding attackers and foreseeable effects. The study made by Im and Baskerville in #2 focuses on human errors, which remain a significant and poorly recognized issue for information system security. The study (2005) proposes and validates a taxonomy of information security threats, which provides additional insight into human error as a significant source of

security risks. For this reason, taxonomies like the one proposed by ENISA take into account threats as unintentional damage, information leakage/sharing due to human errors, or configuration mistakes and errors during software development (unsecure APIs). The research in #3 offers a comprehensive taxonomy of Distributed Denial of Services (DDoS) attacks with the corresponding countermeasures and defence mechanisms. Article #4 proposes taxonomy with four dimensions, which provide a holistic classification, covering network and computer attacks. This taxonomy provides assistance in improving computer and network security as well as consistency in attack description. Article #5 proposes cyber-attack taxonomy and provides nine classes of cyber-attacks. The focus is on effectively classifying attacks. It also proposes taxonomy of cyber adversaries arranged in eight classes according to skill level. In the context of NSF-sponsored work, #6 envisions combining modular threat libraries, i.e. structured or unstructured lists of threats to be coupled with attack taxonomies, which offer a classification scheme to help developers find relevant attacks more easily. The threat list is based on the notion of a threat pattern, which can be customized and instantiated in different architectural contexts to define specific threats to a system. The approach includes a method to construct pattern-based threat taxonomies for more specific system types and/or technology contexts by specializing one or more threat patterns.

ENISA

European Union Agency for Network
and Information Security
Science and Technology Park of Crete (ITE)
Vassilika Vouton, 700 13, Heraklion, Greece

Athens Office

1 Vass. Sofias & Meg. Alexandrou
Marousi 151 24, Athens, Greece

PO Box 1309, 710 01 Heraklion, Greece
Tel: +30 28 14 40 9710
info@enisa.europa.eu
www.enisa.europa.eu