University of Macedonia

Program of Postgraduate Studies

Department of Applied Informatics

Master Thesis

Big Data Real - Time Security Analytics

by

Neofytos A. Kountardas

Submitted as a prerequisite in fulfillment of the partial requirements for the acquisition of the postgraduate degree in **Applied Informatics**

Thessaloniki - Greece, 06/2017

Acknowledgements

This research is my final deliverable of my Master Thesis in order to obtain my Master Degree of Applied Informatics from the University of Macedonia located in Northern Greece. This thesis was a personally strenuous outcome sacrificing precious personal and social time. Therefore, I would admit that my research was a directly individual, yet indirectly family effort, taking into account the alternate costs. So first and foremost, I would like to thank my fiancée that constantly gained me the necessary time to strive for my scientific beliefs while she relentlessly believed in my academic orientation. Secondly, my loving family for their understanding and their long-term support to my personal goals, sometimes by setting aside their real dire needs. Next, I would like to thank my professor of the department of Applied Informatics Psannis E. Konstantinos for his open-minded approach and his constant advisory opinions, while providing me the freedom to drive my thesis according to my vocation and my Security concerns. Finally, I would like to thank my Constitution, namely University of Macedonia and its personnel, from top to bottom, Professors to technical staff, who are really interested in our academic exertion and share our vision for the future.

I am confident that every noteworthy result demands significant effort under limited resources. In my case the driving force behind all, was my firm belief that Security issues in the new digital era demands greater attention and a bunch of moral and well-versed individuals to bring under academic scrutiny all implemented practices regarding Security and Privacy.

Lastly, I really hope my current research to leave a small footprint to the Cyber Security domain and to ultimately add a tiny grain of contribution to the Scientific Community.

Abstract

The magic triangle of IoT, Big Data and Cloud is currently ubiquitous, permeating the digital air around us and pervading into our daily physical and cyber lives. Awareness however is the critical factor when contemplating deploying novelties. Primary goal of this paper is to lay down inescapable security issues and challenges in the new era. Daunting grim thoughts are rendered impotent if we change mindset and utilize the double sword of technological advances in favor of security. What if big data instruments & advanced analytics are deployed selectively to fortify our critical assets from the constant fear of possibly well funded and acutely organized premium attacks? Whether or not the market of security analytics is evolving, the necessity to apprehend advanced security features is a commonplace in the contemporary cyber confrontation. Enterprise editions may have arisen but open source solutions are indeed indispensable. Spanning from Cyber Threat Intelligence and Analytics to recent rapidly developing User Entity Behavioral Analytics, predictive and prescriptive analytics do gain momentum, promising enormous power and numerous security benefits for their users. The already entrenched Hadoop premise has gradually paved the way for advanced distributed confrontation of computationally intensive tasks, however nowadays the trend moves forward to fully capture and demystify the supernatural velocity of generated data in Real – Time, giving birth to real–time optimized decision making. Our Apache Strom deployment was an endeavor to prove that real-time stream processing accompanied with open Security Intelligence feeds can be utilized to enhance our Security countermeasures. Numerous applications of our approach are possible in order to complement a wider defense-at-depth security model.

Keywords: IoT, Big Data, Cloud Computing, Big Data Analytics, Cyber Threat Intelligence and Cyber Threat Analytics, User Entity Behavioral Analytics, Real – Time Security Analytics, Apache Storm

Table of Contents

List of Figures

1 Introduction

1.1 Problem Statement

Nowadays, you can hardly read a technological article without stumble upon at least one term of the magic triangle of IoT, Big Data and Cloud. Terms interconnected, sometimes misunderstood and used interchangeably but easily readable and quite memorable. However novel they may seem, there are been around for some years now, growing gradually their reputation among vendors, scientists, students and final users. Unfortunately not only proponents exist, but primarily it is vital to accept that these terms are not a current trend or just a temporary topic for ICT practitioners or theoreticians, but an already mature and deeply entrenched fundamental part of our daily diverse life, touching almost every scientific domain independently of its nature or origin. From our inner physical organs (implantable nano-sensors) to our external supernatural frontiers of the universe (astronomical measurements in YottaBytes) common features and characteristics can be found, thus stored, disseminated, processed and thoroughly analyzed, unfolding hidden patterns and giving unprecedented insights and ultimate intelligence for constant real-time optimum decision making and self-regulation. But let's take things from the beginning.

Even though Internet of Things – IoT, was first coined by Kevin Ashton in 1999 (Ashton, 2009) in the context of supply chain management, many definitions can be found in the literature with the prevailing thorough explanation of R. van Kranenburg in 2008, where IoT is "a dynamic global network infrastructure with self-configuring capabilities based on standard and interoperable communication protocols where physical and virtual 'Things' have identities, physical attributes, and virtual personalities and use intelligent interfaces, and are seamlessly integrated into the information network" (Kranenburg, 2008).

Big data is considered to be "a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex and of a massive scale" (Hashem et al., 2015). Big data is currently defined though 4V's, namely volume, variety, velocity and veracity (Gantz et al., 2011) (Zikopoulos et al, 2012), even though 5V's references do exist incorporating the fifth V of value (Terzi et al. 2015). Volume accounts for their enormous size measured in Pettabytes and beyond, velocity for their highly generating speed, variety for their disturbing heterogeneity, veracity for their questionable trustworthiness, and value for their potential gain or pain. Moreover, Big data are classified based on five aspects: (i) data sources, (ii) content format, (iii) data stores, (iv) data staging, and (v) data processing, which are explained briefly in Figure 1 and in more detail in (Hashem et al., 2015).

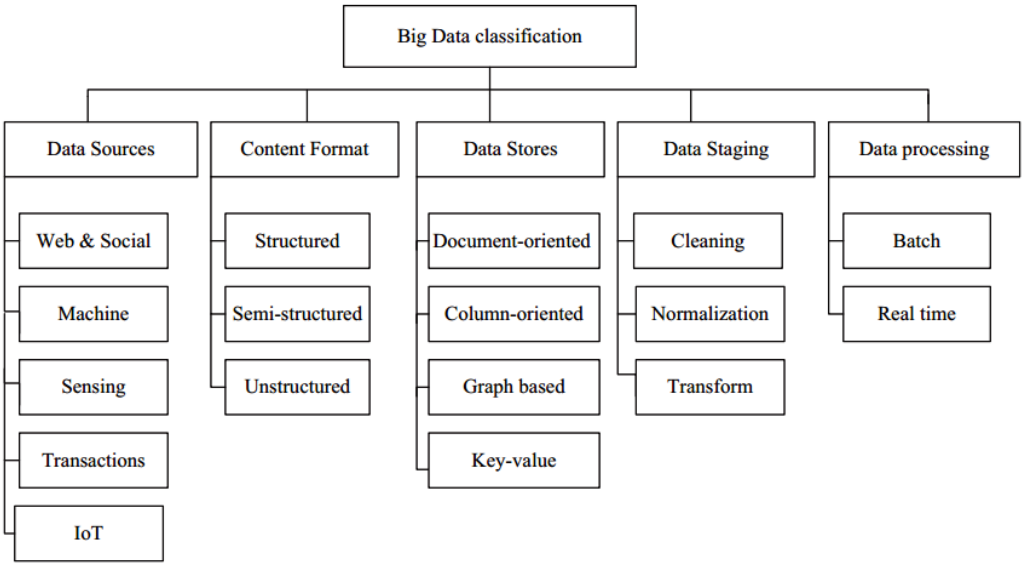


Figure 1. Big Data Classification (Hashem et al., 2015)

In our modern life, Big Data are incessantly generated from an extremely wide variety of data sources, like sensors and actuators, RFID tags & QRcodes, home appliances, local or global networks, logistics, grids, every kind of transaction, emails, documents, videos, audios, images, financial records (stocks, bonds or coins), click streams, logs, posts, search queries, personal files & health records, social networking interactions, scientific data, smartphones & applications data, 3D models, graphs, simulations, Unmanned Aerial/Maritime/Underwater Vehicles (UAV/UMV/UUV), GPS devices and geolocation information and unimaginably boundless more sources. Although Big Data are ubiquitous with new extraordinary processing, storage and manipulation requirements, they are simultaneously promising enhanced decision making, hidden patterns and insight discovery, process optimization and many more smart solutions to each and every aspect of our physical and cyber life. Not to mention the broad business value when Big Data are efficiently combined with traditional structured data (relational databases) from conventional business applications, such as Enterprise Resource Planning (ERP) and Customer Relationship Management (CRM).

As figure 2 illustrates, another way to characterize the big data domains is to look at the types of industries that generate and need to extract information from the data (CSA, 2014).

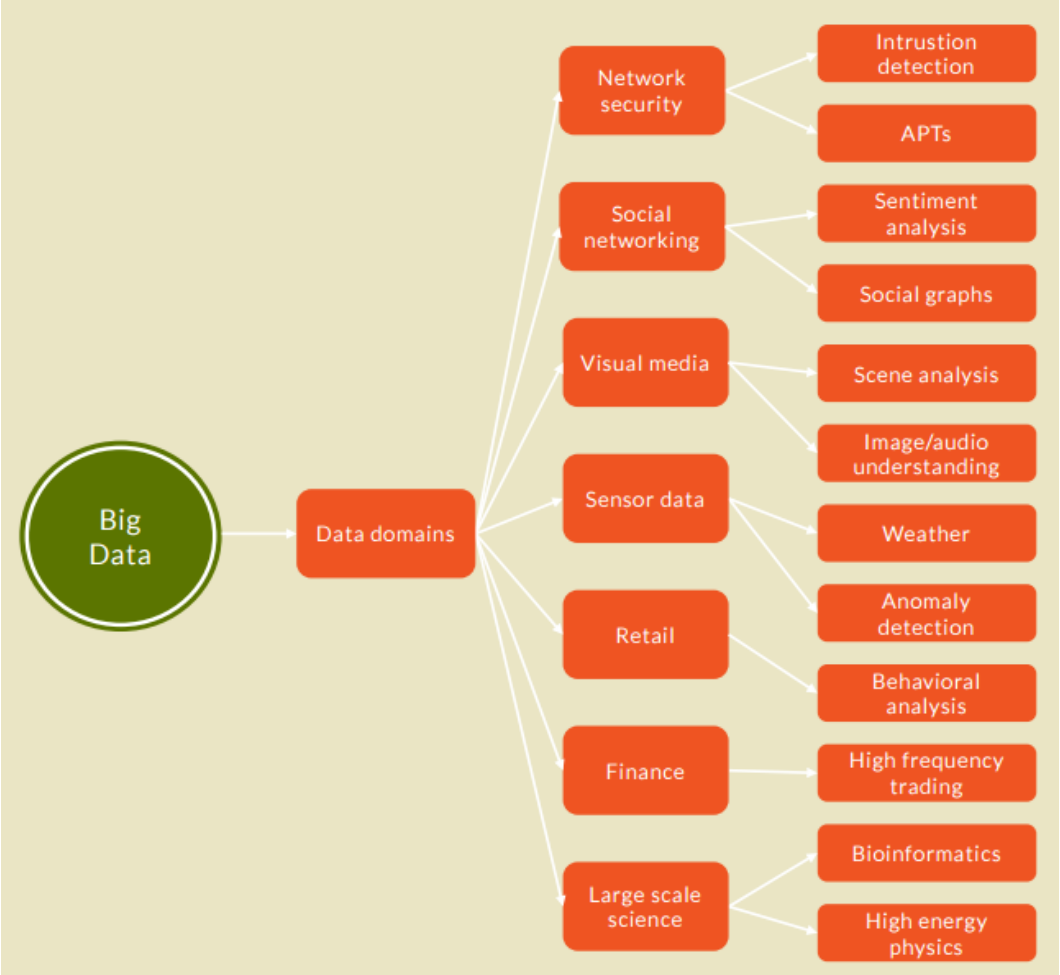


Figure 2. Data domains (CSA, 2014)

Furthermore, CSA illustrates in figure 3, how the big data verticals map to the time and organization axes. Accepting the case that all included industries have use cases that encounter data at all organizational levels and processing needs that span all response times, the industry domain is another orthogonal axis for characterizing the domain space of big data. So, these domains were visualized by particular common use cases, and were mapped to industry, time, and structure (CSA, 2014)..

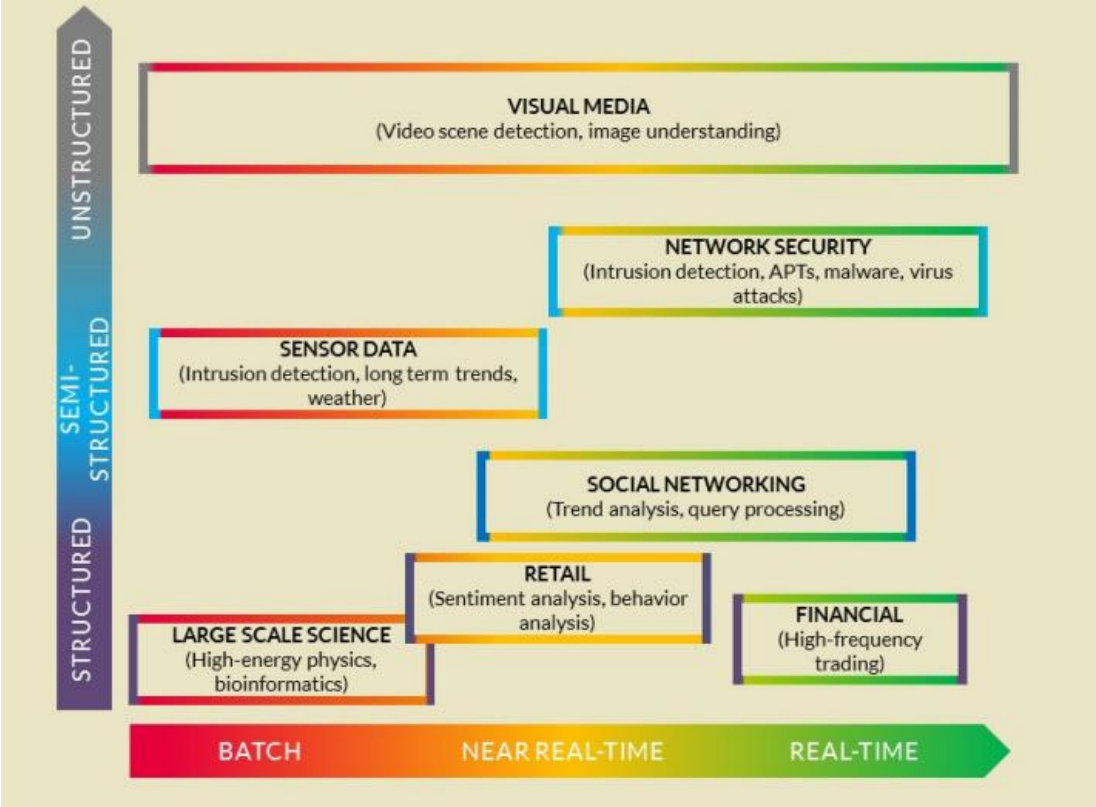


Figure 3. Big data verticals (CSA, 2014)

Contemplating that in **every second** we have **2.506.278** Google searches, **16.111** Tweets, **1.375.000** YouTube video views, **2.600.715** emails sent (67% of which is spam!), almost **45TB** of internet traffic, **14.444** photos uploaded to Instagram and **2,643** calls made via Skype, all along with the fact that every minute **300 hours** of video is uploaded to Youtube and almost **5 billion** videos are watched every single day or that every hour **9 million** messages are sent on Facebook the volume of data is just amazing (Statistic Brain, 2017) & (Internet Live Stats, 2017). We should also consider that "The amount of data generated is expected to double every two years, from 2500 exabytes in 2012 to 40,000 exabytes in 2020"(CSA, 2016).

Our world is moving extremely fast becoming each second more digital in an unprecedented rate. In this environment, Big Data with their diverse, complex, and massive-scale nature have rendered useless our traditional computing techniques and a smooth transition path from traditional to big data was inevitable. Therefore, new methodologies have gradually emerged paving the way to a more sophisticated utilization of really big data sets, releasing enormous power and unforeseeable knowledge.

A remarkable example of Big Data is the aviation industry. As Senior Director of Marketing, Aerospace and Defense at SAP Thomas Pohl explains, data collected by sensors on one aircraft usually covers more than 300,000 parameters (Thomas, P. 2015). Out of these parameters engine data are one of the most important set of data points they capture. Considering an average commercial aircraft (i.e. Boeing

737) he mentions that as soon as an aircraft has two engines, each engine creates 20 terabytes of information per engine hour, combined with an average six-hour cross-country flight from New York to Los Angles yields 240TB of data for every engine hour. Only in the United States there are approx. 28,500+ commercial flights in the sky on any given day, multiplied with 365 days a year and we have reached a real big data challenge. Not to mention that this is just for commercial flights and merely the data from two engines. We could also take into account the US National Air Traffic Controllers Association which estimates a total of 87,000 flights in the skies in the United States on any given day. Nevertheless, Pohl claims that the number of engine data collected only from the commercial aircraft engines in one year for the United States is 2,499,841,200 TB (figure 111), in other words to approximately 2,7 Zettabytes which equals nearly three times of the total estimated global IP data traffic per year in 2015.

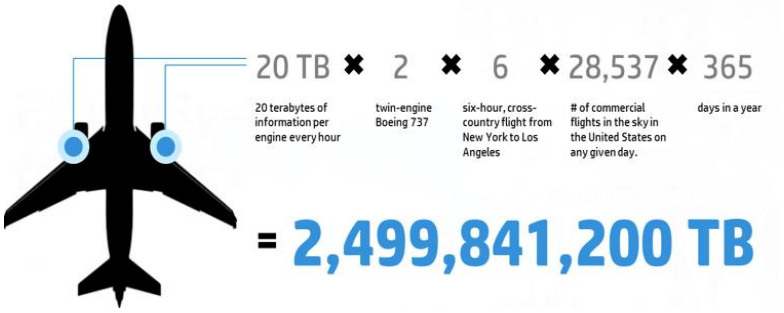


Figure 4. Sensor data from a cross-country flight (Thomas, P. 2015)

Such Big data figures are useless, unless cost impact and revenue potential is estimated. Probably, in the aviation industry sensor data are frantically obtained to maximize operational efficiency, by optimizing maintenance processes and other related issues to gain more "flying time" for their fleet. Pohl states "when an airline operator acquires a $100-390 million aircraft for its fleet, the goal is to keep the aircraft up and running for at least 18 hours a day for the next 15-20 years. According to International Air Transport Association (IATA) the leading cause of late flights (42%) are based on airline-controlled processes, such as maintenance. Every hour the aircraft is not in operation, costs the airline operator an avg. $10,000". Therefore "Predictive analysis can help turn huge amounts of maintenance-relevant data into actionable information, helping ensure that maintenance technicians execute the right work steps at the right time and with the right tools. Predictive analysis can help drive strategic improvements and provide better-quality output at lower operating cost and improved return on investment".

Next came Cloud computing – (CC). CC is well defined as "a model for allowing ubiquitous, convenient, and on-demand network access to a number of configured computing resources (e.g., networks, server, storage, application, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction" (Mell et al., 2011). CC is widely accepted as the next generation of IT, promising reliable on demand software (SaaS), hardware (IaaS), and applications services

though platforms (PaaS) delivered over the Internet, while services on cloud generally propose powerful architecture to perform complex large-scale computing tasks. The unmet need to store, manipulate and analyze massive amounts of data has driven many organizations and individuals to adopt CC (Liu, 2013), while extensive scientific experiments are currently deployed in the cloud and may continue to increase because of the lack of available computing facilities in local servers, reduced capital costs, and increasing volume of data produced and consumed by the experiments (Pandey et al 2013). The increasing popularity of wireless networks and mobile devices has taken cloud computing to new heights because of the limited processing capability, storage capacity, and battery lifetime of each device (Khan et al., 2014).

After that, it becomes apparent that CC is perfectly integrated with IoT and Big Data in a unique matching manner. The onslaught of data generated of each and every tiny to supercomputer device is aggregated and consolidated via CC architecture and possibilities. Therefore a supreme triangle among them is created.

It is necessary to clarify that in this thesis, big data references usually encompass abstractly the massive amount of data generated by various interconnected devices, all along with the architecture of CC implementations and its participation in extensive storage, dynamic manipulation and wide transmission of all generated data as well as other enabling technologies conducive to the interconnection of information.

1.2 Motivation and Research Goal - Contribution

All in all, the triangle of IoT, Big Data and Cloud Computing is extremely compelling and promising. But should we pose a question about Security? Do these novel paradigms enhance or hinder our Privacy? It is controversial to speak about privacy in an era which every single move could be captured, filed, transferred, aggregated, automatically processed and analyzed yielding results and subsequently decisions about human behaviors, sometimes proceeding in further automatic acts hopefully accurate and primarily ethical in a massive and difficult controllable obscure way. This serious topic has sparked unrelenting debates over IoT, Big Data or Cloud implementations and raised dire concerns about the potential harms.

Therefore our motivation is to shed some light to the Cyber Security domain, to outline dreaded Security concerns and open issues and to comprehend the current development stage and ongoing lifecycle phase, especially regarding novel technologies and paradigms.

The basic idea dispersed into this research was a kind of speculation on whether novel technologies pose only Security issues or can be grasped and utilized for the sake of Security; a personally interesting and motivational aspect of current technological advancements.

Furthermore, we present the core concepts behind a famous Real – time processing framework (Apache Storm) tweaking its units and developing a tiny application for enhanced Security, by processing real –time big data. To this end, evaluating our conclusions, we can assert that indeed novel paradigms can and should be implemented to our battle against crimes or just disturbing harms.

We hope that our contribution exceeds the current limits of real –time processing for enhanced Security while informing our readers for Security advancements and open issues.

1.3 Thesis Overview

Presenting a solid security framework seemed compulsory in this article, in order to clarify basic concepts among security issues, so chapter 3 is dedicated to that, followed by predictions about cyber security in the near future and the reason that security matters. Afterwards, security challenges are listed in the next section 4, with well-known organizations' reports and scientific articles on these issues along with recent notorious attacks revealing serious vulnerabilities, lastly in this chapter, we questioned the notion of big data, focusing on whether we can harness the potential power of new big data analytics techniques and methodologies for a safe and secure cyber environment. However, this paper is not trying to address or remedy theoretically any of these significant issues. We know that currently there is enough literature on this aspect, yet we hope that this specific domain will be continuously under examination by many forthcoming researchers. On the contrary, this paper mainly doubted about the single-side evil nature of big data, therefore we flipped that coin and examined the area of big data for an enhanced security. In order to do so, in section 5 basic notions on advanced security analytics and intelligence were presented, clarifying the landscape of intertwined security terms. Later on, chapter 6 explained our methodology and our approach that is steadily unfolded throughout this thesis. Acknowledging that big data concepts will have crucial yet promising impact in Cyber Security Intelligence, in section 7 we introduced big data alignment with security. However, recent advances in field promoting enhanced security analytics are under scrutiny especially in chapter 8, where in brief security analytics platform's characteristics are depicted. Big data are distinguished from traditional data in various ways and this offers many advantages to security academic or organizational seekers. In chapter 9 we enumerated some open source ideas for enhanced security analytics, but we elaborated on revolutionary Hadoop ecosystem in chapter 10, considering the differences between batch and real-time processing. Afterwards, and in a lambda architecture approach we opted for Apache Storm. So, in chapter 11 after describing Storm's internal parts and impressive characteristics, we backed our decision considering storm a powerful tool for reliable and scalable real – time processing. Our thesis ends up with a major practical implementation of Stream processing in chapter 12, using basically Apache Storm, combined with tailor-made automatically updated Security Intelligence

feeds, utilizing many open source big data tools, concepts and methodologies. Primarily our application was created for dropping packets from unwanted and potential malicious users of our web server, based on blacklisted IPs. However, many example applications were enumerated in sub-chapters of chapter 12, outlining in parallel their limitations. Finally, in chapter 13, conclusions will terminate this research creating the foundation of further scientific research upon advanced cyber intelligence and analytics platforms.

2 Literature Review

Regarding "hot" issues like Big Data, IoT and Cloud computing, the public interest has its peaks and troughs. Similarly, academia and business follow a common pattern when referring to technological advancements and emerging promising technologies. Gartner hype cycle, is a unique manifestation of this interest at emerging technologies, and partial components (IoT platform, Connected Home, Personal analytics, UAVs) of our big comprehensive notions, like big data and IoT are depicted in this hype cycle.

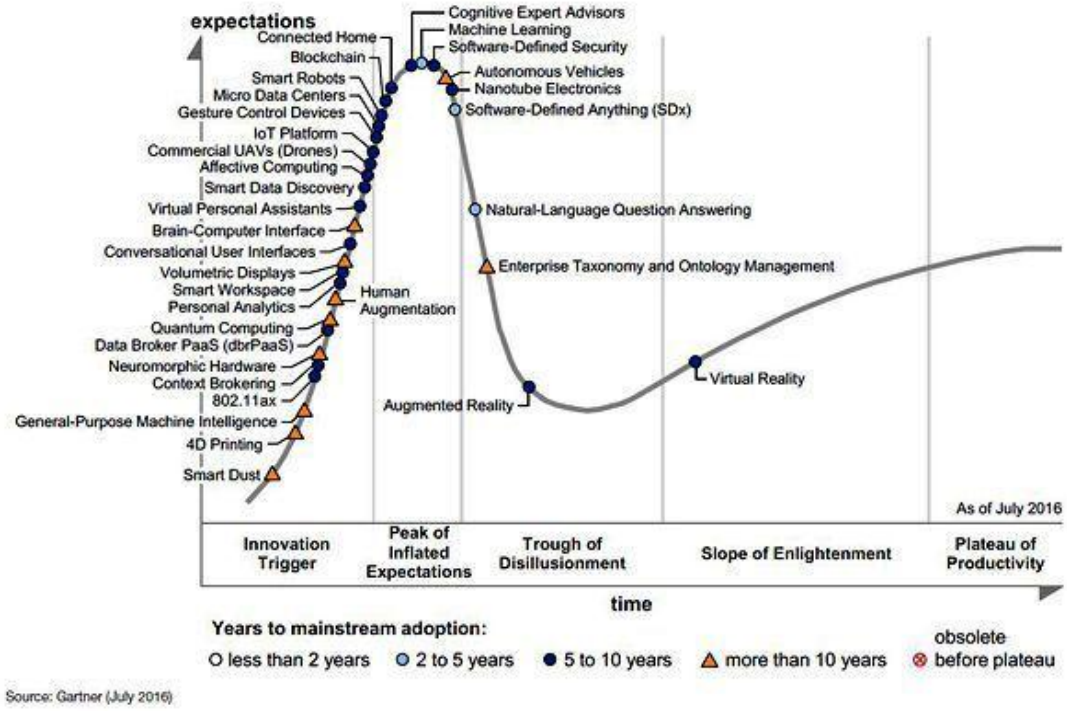


Figure 5. Hype Cycle for Emerging Technologies (Gartner, 2016)

Trying not to be verbose about Gartner's annual insights – analysis, this hype cycle was used to promote the interesting pattern that most technologies follows, along with their corresponding reviews, literature and public interest.

So, literature about technological issues remains enormous and substantial effort is needed to gain fine-grained, accurate and not outdated knowledge. A vast amount of peers are everyday seeking to reveal optimized and/or novel paradigms that outperform the current ones.

In our thesis, we considered proper to mix all possible resources with the only ultimate goal to attain unbiased and fully objective knowledge. Therefore, academic scientific papers, business' products and services, reviews or white papers, researchers' publications, public and private institution white papers, recommendations or guidelines, corporate blogs, project documentations and many more resources were aggregated to pave the way for this thesis, in order to draw the big picture of our chosen notions.

Apparently, when inundated with too many resources we prioritized contemporary and mostly cited references, as well as entrenched and well-versed corporations.

3 A clear reference model about Security

Approaching the domain of Information Security could be a daunting task, unless a solid framework was deployed in order to remain on rails across a broad scope of abstract notions and complicate definitions between almost identical but deeply different concepts. Therefore, we opt for the Reference Model of Information Assurance & Security - (RMIAS) (Cherdantseva et al, 2013), wherein researchers after analyzing all the necessary terminologies involved in Security (Information Security - Infosec, Information Assurance - IA, Information Assurance & Security – IAS, Reference Model - RM, Information System – IS), explained how open interconnected environments (de-perimeterisation), intensive collaboration, e-commerce, outsourcing and cloud computing have rendered IAS more challenging than ever, therefore they proposed their reference model, the so-called RMIAS, in order to overcome the limitations of contemporary existed conceptual models of InfoSec and IA. Building upon CIA-triad (confidentiality, integrity and availability) originated back in 1975 (Saltzer et al., 1975), they covered all intermediate steps of IAS development. Unfortunately, the lack of a clear set of security goals, the absence of a high level of abstraction classification of security countermeasures and the generic consideration of time among other security models, urged them to articulate their RMIAS.

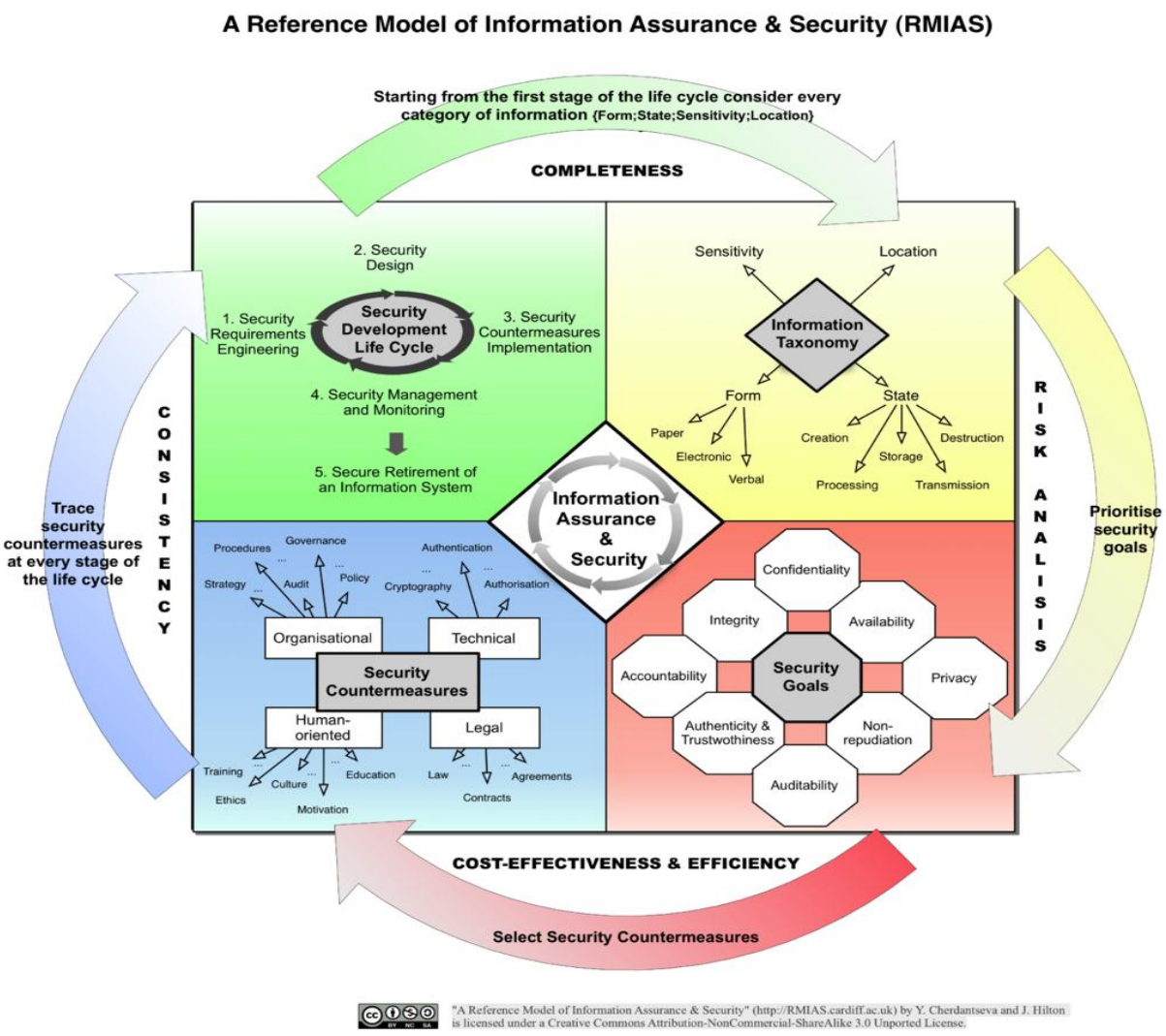


Figure 6. RMIAS

In a holistic approach of IAS domain, the RMIAS, depicted in Figure 01, has four major dimensions: 1) Information System Security Life Cycle Dimension (Security Requirements Engineering, Security Design, Security Countermeasures Implementation, Security Management and Monitoring and Secure Retirement of an Information System), 2) Information Taxonomy Dimension (Sensitivity, Location, Form like paper, electronic or verbal, State like creation, processing, storage, transmission and destruction), 3) Security Goals Dimension (Confidentiality, Availability, Privacy, Non-Repudiation, Auditability, Authenticity & Trustworthiness, Accountability and Integrity) and 4) Security Countermeasures Dimension (Organizational; strategy, procedures, audit, governance and policy, Technical; cryptography, authentication and authorization, Legal; law, contracts and agreements, and Human-Oriented; training, ethics, culture, motivation and education), with interrelationships between them in cycle manner. The above visual representation of IAS endeavors to organize the security domain knowledge and to make this model memorable as well as easily understandable and widely referenced.

It is noteworthy to mention that Cherdantseva et al, in their second orchestrated approach, came back with a straightforward goal; to verify the following hypothesis: "*The RMIAS provides more complete and accurate representation of the IAS domain, than the existing conceptual models of the IAS domain. The RMIAS reflects how the IAS domain is understood by IAS domain experts. It represents the domain in the form accessible by the experts with the different backgrounds and with the different levels of experience in IAS. Due to the above, the RMIAS helps to build a congruent understanding of the IAS domain in a multidisciplinary team of experts".* Along their remarkable multifaceted and multi-criteria methodology, the aforementioned hypothesis is ultimately triumphantly verified through an extensive analytical and empirical objective evaluation, which included twenty-six interviews with IAS experts, three workshops and a case study. Finally, the RMIAS was characterized as a suitable cognitive model and a basis for building a congruent understanding of the IAS domain in a multidisciplinary team of security and non-security experts (Cherdantseva et al, 2016).

It is intuitive to infer that from that angle, the whole premise of security is tremendous and could incorporate too many interrelated aspects. Consequently in this paper we have zeroed in only specific fields from this wide approach, focusing basically on technical security countermeasures through real – time big data analytics taking into account the security development life cycle for the real – time processing of information with the ultimate goal of securitize critical assets and maintain Privacy, Confidentiality and Integrity in our perimeter. Our first and foremost however goal remains the "Security Awareness" and the fact that knowledge around the domain of Information Security is more than necessary, therefore our paper demands attention and lay down security facts with an straightforward educational motive.

3.1 Why Security matters

One could consider Security issues as an outdated topic out of interesting scope, on the grounds that extensive deployments of pervasive and ubiquitous computing are self-proving their safe and secure nature. Unfortunately, even though Smartphones (and other vital appliances) have become precious inseparable personal belongings resembling a kind of assistant-living, nothing is securely addressed so far. Contrary to naive beliefs, significant security concerns permeate the cyber world and dark thoughts are raised when the news transmits rumors about malicious Russian cyber intervention into US elections and other governmental institutions (Harding, 2017).

In the brick of 2017, almost every day, organizations are inundated with numerous reports about major cyber-incidents, either unsettling when referring to their own critical infrastructure and assets, or much to their relief whenever concerning another rival or adversary. However, these reports are directly or

indirectly hinting at a constant fear of a cyber attack that could render the whole organization inactive and damage it irreversibly.

Moreover, PwC in its Global Economic Crime Survey for 2016 (PwC, 2017), revealed that Cybercrime is the second most reported economic crime after misappropriation, while 32% of organizations that PwC examined had already been affected by cybercrime. Also, over 1.000.000$ losses were reported for 14% of organizations in question, and notably the 1% of all lost over 100m$.

Eventually, large corporations reallocate their budget taking seriously into account security requirements. Although security spending is at an all time high, security breaches are also at an all time high, according to Gartner Inc (Rivera, J. & Meulen R., 2015). The impact of advanced attacks has reached boardroom level attention, and this has freed up funds against such attacks.

Looking backwards to the near past cyber-attacks along with the current trends of Computer Science, predictions can be made about Cyber Security in the near future. As WhiteHat Security reports (Zubair, A. 2016), high profile data breaches are currently on the rise. Well-known organizations, including government agencies, fell victims to cyber-attacks, but fortunately, data breaches have increased awareness and forced organizations to advance their security practices to better fight cyber-threats. WhiteHat believe that these trends will continue into 2017, listing top security predictions for 2017, specifically: 1) Applications offer a large, vulnerable attack surface to hackers and will continue to be the weakest link. According to the 2016 Verizon data breach report (Verizon, 2016), 40% of data breaches come from web application attacks. Meanwhile WhiteHat's Statistics Report (Hardof T., 2016), states that application security is not pretty, remediation rates are under 50%, and vulnerabilities that are eventually remediated stay open for months, without any foreseeable fundamental change in app development processes or security practices. As a result, big data breaches originating from the application layer are anticipated to continue to hit the headlines in 2017. 2) IoT attacks will continue, because billions of IoT devices currently in use are insecure, and are not likely to be patched or fixed anytime soon. Consequently, hackers will continue to use insecure IoT devices to launch attacks like the one on Dyn in Oct, 2016 (Hilton, S., 2016). New government regulations are expected in this area to require manufacturers to build better security in IoT devices. 3) Cyberattacks stemming from weak, stolen or default credentials will decline. According to the 2016 Verizon data breach report (Verizon, 2016), 63% of confirmed data breaches came from default, weak or stolen passwords. Fortunately, multi-factor authentication – (MFA) is gaining popularity, according to the research firm Markets and Markets which estimates that the MFA market will reach 9.6 Billion USD (a CAGR of 17.7%) (Markets and Markets, 2016), expecting that the growing adoption of MFA will curtail this kind of cyber-attacks. 4) Vendor security risk will become more manageable. Vendor security risk, a known area of concern due to ad-hoc processes and lack of transparency from vendors, is

undergoing transformation, on the grounds that new sections were added to PCI standards to clarify the responsibilities of service providers in the PCI compliance process. Moreover, the Vendor Security Alliance - (VSA), a coalition of companies committed to improving Internet security by streamlining vendor security compliance, released on October 1st 2016 its first (annual) questionnaire, helping organizations to assess and benchmark third party product and service risk, while ensuring that appropriate controls conducive to improved security are in place. After that, vendor security risk management processes are expected to be more streamlined and automated. 5) Organizational silos will give way to a more security-centric culture. Separate organizational silos remain the last frontier for enhanced security stemming from different prioritization or unwillingness to collaborate. The development team's priorities are often not aligned with security team's priorities, and even within the security team, folks working on Application Security don't necessarily collaborate with network security or cloud security teams. Fortunately, organizations are streamlining processes and aligning priorities. The widespread optimism of WhiteHat lies on the expectation that 2017 would be a turning point, moving forward to a greater adoption of DevSecOps, better risk management maturity, greater information transparency and collaboration.

As David Burg (US and Global Co-Leader in Cybersecurity - PwC) points out: "We're seeing more and more that cybersecurity can actually become a remarkable way to help a company innovate and move faster. In certain kinds of digital innovation, the security considerations, controls and capabilities, alongside a frictionless means of authentication, are essential to the design and development of these new products and services."

4 Security Issues & Challenges in IoT & Big Data

As Cloud Security Alliance - CSA has remarked in (CSA, 2013), security and privacy issues are magnified by the volume, variety, and velocity of big data. Large-scale cloud infrastructures, diversity of data sources and formats, the streaming nature of data acquisition and high-volume, inter-cloud migration, all play a role in the creation of unique security vulnerabilities. Today, big data is cheaply and easily accessible through public cloud infrastructure and organizations of all sizes have access to the information and the means to employ it, contrary to the past, when big data was limited to very large users (governments and sizeable enterprises) that could afford the necessary infrastructure for hosting and mining large amounts of information. Furthermore, Software infrastructures like Hadoop which enable deploying thousands of computing nodes to perform data-parallel computing, have accelerated adoption of big data mining methodologies As a result, new security challenges have arisen and as big data are expanding into streaming cloud technology with unprecedented demanding ultra-fast response times from

security solutions, traditional security mechanisms tailored to secure small-scale, static data on firewalled and semi-isolated networks are inadequate.

Taking these into account CSA highlighted the top ten Big Data security and privacy challenges that need to be addressed to make Big Data processing and computing infrastructure more secure. These challenges are enumerated as: 1) Secure computations in distributed programming frameworks, 2) Security best practices for non-relational data stores, 3) Secure data storage and transactions logs, 4) End-point input validation/filtering, 5) Real-time security monitoring, 6) Scalable and composable privacy-preserving data mining and analytics, 7) Cryptographically enforced data centric security, 8) Granular access control, 9) Granular audits and 10) Data provenance. The following figure depicts the top ten security challenges.

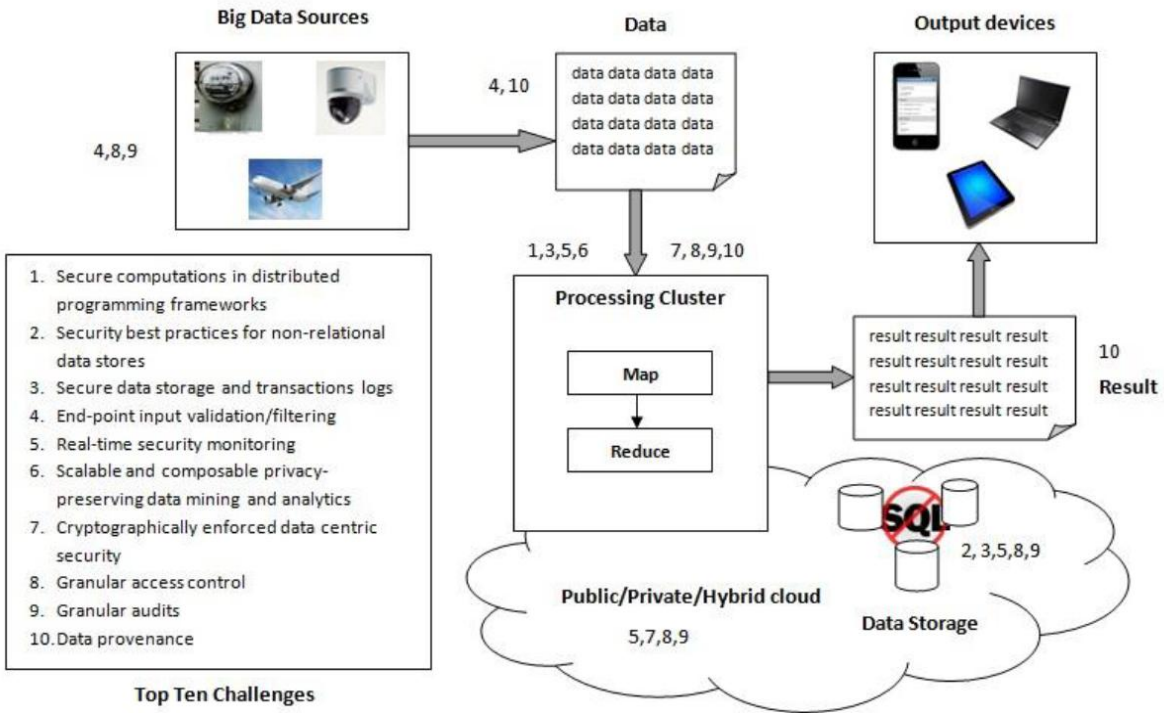


Figure 7. CSA's Top 10 Security & Privacy Challenges in Big Data Ecosystem (CSA 2013, 2016)

Subsequently, CSA organized all related challenges into four aspects of the Big Data ecosystem, creating four discrete areas, depicted in figure 03. These aspects are: 1) Infrastructure Security, 2) Data Privacy, 3) Data Management and 4) Integrity and Reactive Security infrastructure. Furthermore, CSA in the same report provides a brief description of each challenge, reviews usage of Big Data that may be vulnerable and summarizes existing knowledge according to the modeling, analysis and implementation approach for each challenge.

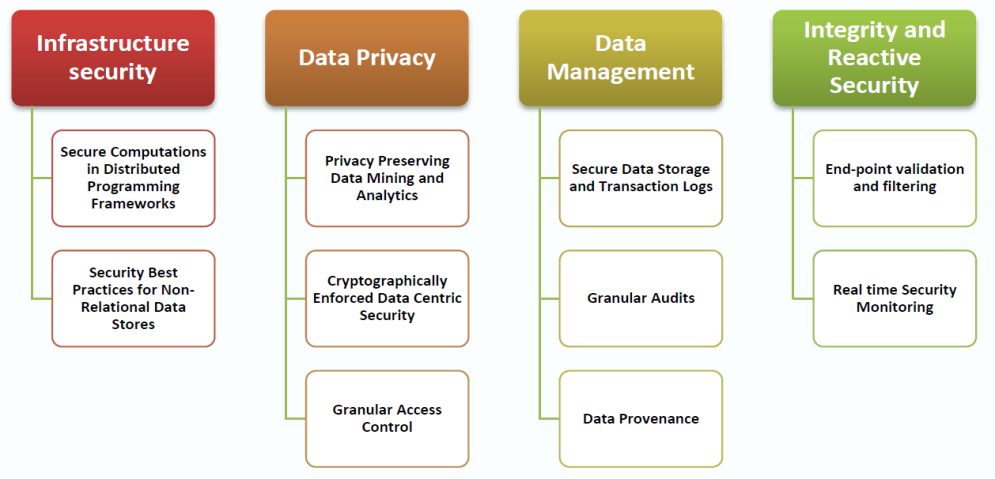


Figure 8. CSA's classification of the top 10 challenges (CSA 2013, 2016)

Additionally, European Union Agency for Network and Information Security - ENISA has identified (Naydenov et al, 2015) the following key challenges related to the secure use of Big Data: 1) Access control and authentication, because maintaining the desired level of access control and authentication of all participating entities is a potential problem, 2) Secure data management, on the grounds that the big volume of logs needs to be stored appropriately, protected properly and if need be securely restored, 3) Source validation and filtering, because verified and trusted sources can guarantee more accurate results. Moreover, recommendations have been issued for participating stakeholders, thus: 1) Policy makers should focus on providing guidance for secure use of Big Data systems in critical sectors, 2) Standardization bodies should adapt existing standards or create new security standards including Big Data, 3) Big Data providers or vendors should comply with security standards, 4) Competent authorities should encourage vendors to offer security authentication mechanisms, 5) Industry players and vendors should invest more into staff trainings and certifications enhancing their technical security skills.

Terzi et al. (2015) approached big data ecosystem from its privacy and security perspective, accumulating all related literature. They categorized the big data security and privacy issues under five titles, named Hadoop security, cloud security, monitoring and auditing, key management and anonymization and afterwards they examined and classified big data security and privacy studies into aforementioned categories. Moreover, they discussed the role of big data analytics for enhanced security, by exploiting the increase of stored and streamed data and development of analysis systems. Accumulating large volume and variety of data and associating them with network history can expand current anomaly, intrusion, fraud and advanced persistent threats detection systems.

The Cloud Security Alliance (CSA) has recently listed out, the best practices that should be followed by big data service providers to fortify their infrastructures (CSA, 2016), presenting 10 considerations for each of the top 10 major challenges in big data security and privacy. Totally 100 best

practices are presented, structured in way that answers: What is the best practice? Why should it be followed? (Giving examples of specific security/privacy threat at stake), How can the best practice be implemented?

Thuraisingham (2015) mentions serious security and privacy considerations when referring to big data. On one hand, raising concerns on combinations of disparate data about an individual, on the grounds that even when data are stripped of personally identifiable information (PII) there are techniques to create uniqueness and identify individuals and on the other hand pointing various regulations which may cause privacy violation or just stifle innovation. Therefore a balanced approach towards regulations and analytics is needed. Privacy preserving techniques for data mining, data integration and information retrieval are enumerated while collocating the inherent inadequate security protections with Hadoop, MapReduce, Hive, Cassandra, Pig Latin, Mahout and Storm. Finally, contrary to the aforementioned security issues, the potential of big data analytics regarding enhanced security is acknowledged. Specifying examples of organizations that could outsource activities such as identity management, email filtering and intrusion detection to a central cloud and this massive amount of data collected could be analyzed for security applications.

Lafuente (2015) considered real-time security monitoring as key security component for a big data project. It is important that organizations monitor access to big data to ensure that unauthorized access is being carried out. It is also vital a threat intelligence system to be in place to ensure that more sophisticated attacks are detected and organizations can react to known security threats accordingly (malware, vulnerabilities and bugs).

Addressing privacy risks from the very beginning of the processing and applying all necessary privacy preserving solutions in every stage of the big data value chain is considered “Privacy by design” by the European Union Agency for Network and Information Security – ENISA (Acquisto et al, 2015). Specifically, ENISA endeavors to find all appropriate mechanisms to implement privacy principles in the big data environment, and to this end, the concept of privacy and data protection by design remains fundamental. Privacy by design can empower individuals in the big data era as well as support data controllers’ liability and trust. This is why reason ENISA presented an overview of privacy by design in the big data value chain, analyzing concepts and design strategies of the “privacy by design” notion and exploring challenges in practical implementations. Finally some privacy enhancing technologies are proposed, focusing on anonymization, encryption, accountability and accessibility, transparency and consent.

4.1 Security and Privacy challenges in IoT

One promising application of IoT is home automation, with smart home technologies like sensors and actuators, energy controllers, smart appliances and many more conveniences, spanning from automatically adjusted thermostats and smart door locks (i.e. using geofence) to garage doors (synced with mobiles reporting their open or closed state). However, enormous security and privacy issues are raised, when contemplating an adversary with access to smart home sensor data, detecting owner's moves and habits. To this end security and privacy problems have been encountered across devices and networks, from falsified medical devices (pacemakers), networked light bulbs (insecure transmitters and receivers – mesh networks), Nest thermostats, fitness trackers (FitBits), toasters to smart TVs that listen to conversations (BBC, 2015) and refrigerators enlisted into denial of service attacks (Newman, 2014).

In smarthome environment, all home devices are accessible via Internet anywhere and anytime with the ultimate goal to convenient our daily lives. However the more home devices are linked, the more security flaws are revealed. Many security experts warned that in house appliances could be used as attacking tools. Unfortunately smart TVs and refrigerators infected by malicious code capable of sending bulk spam emails is a reality. If we consider that basically Smarthome consists of 4 parts (smart devices, home network, home gateway & service platform), in this paper is stated that service platform which is normally located in service providers, could be used for delivering various services to home devices, taking security into account initially. Specifically, for each part security vulnerabilities that could be met are listed, so in smart devices DDoS, information leakage and unauthorized use, in home gateways DDoS, information leakage, in home network unauthorized access, eavesdropping or falsification and finally in service platform eavesdropping or falsification (Yoon et al, 2015)

Schurgot et al, (2015) are exploring the risks of security and privacy in IoT networks, focusing first on home automation networks and on mechanisms for privacy preservation in smart homes generally. In particular, they experiment with techniques to trade-off resilient use of IoT-enabled services with protection of user data privacy, seeking user protection at both simple cryptographic techniques and information manipulation when an adversary has compromised remote servers or has just gained access inside the IoT network. Their goal is to experiment with techniques that manipulate user's IoT data in order to mislead adversary's view. In this way the user could reconstruct the true state of their data while the adversary has uncertainty about the true state, increasing the difficulty in conducting a successful attack.

A notorious example of IoT vulnerabilities and the new era of unprecedented massive attacks exploiting new technologies is the recent attack to Dyn Company (Hilton, S., 2016). Dyn which controls massive DNS infrastructure was hit on 21 October 2016 and remained under attack for most of that day, bringing down sites including Twitter, the Guardian, Netflix, Reddit, CNN and many others in Europe and

the US. A distributed denial of service (DDoS) attack, by a "Mirai" botnet, collapsed Dyn DNS Servers and unlike other botnets, which are typically made up of computers, the Mirai botnet was largely made up of IoT devices. Dyn estimated that the attack had involved 100.000 malicious endpoints, and the company said there had been reports of an extraordinary attack with the strength of 1.2Tbps. If those reports are true, that would make the 21 October attack roughly twice as powerful as any similar attack on record, converting it into the largest DDoS attack on record. Mirai was also used in a former attack on the information security blog "Krebs on Security" (Krebs. B., 2016), run by the former Washington Post journalist Brian Krebs, in September 2016 at the lower rate of 665 Gbps. David Fidler, adjunct senior fellow for cybersecurity at the Council on Foreign Relations, said "We have a serious problem with the cyber insecurity of IoT devices and no real strategy to combat it".

4.1.1 IoT, Big data and Cloud Computing combined

Big Data has common and overlapping areas with IoT and (Mobile) Cloud Computing (MCC), therefore unique features mutually beneficial do exist, in a wide spectrum Security inclusive. Evidently, the technologies of Mobile Cloud Computing and IoT began their developing basically due to the technology of wireless networks. In (Stergiou, C., & Psannis, K. E., 2017) MCC, IoT and Big Data were combined, in order to check the common features and discover the benefits of MCC and IoT when used with Big Data applications. In their paper, a survey of IoT Technology was presented and main features of MCC were enumerated. Most notably, the contribution of IoT and MCC to Big Data applications was elaborated. As a conclusion, IoT mostly contributed to the field of network partitioning and unified Ethernet fabrics of Big Data applications, while MCC mostly contributed to the field of Predictable and efficient, holistic network, and network partitioning of Big Data applications.

4.2 Sensitive Healthcare areas

IoT Security issues can easily gain wide attention on the grounds of the pervasiveness of this kind of technology. Healthcare is considered a major domain of IoT applications with a variety of use cases impacting the current healthcare programs. The advantageous nature of IoT technology deployed in the Health domain, could generate benefits for patients, health professionals and insurance companies. But shouldn't we be aware of novelties' misconfigurations or security backdoors when referring to ultimate goods like our health? Madelyn Bacon, an assistant editor in SearchSecurity, has recently published (on 13 Jan 2017) an interesting article, by which she has raised significant awareness around the controversial issue of IoT implantable devices. The article was mainly dedicated to St. Jude Medical (SJM), a till recently American global medical device company headquartered in Little Canada, Minnesota, U.S., which builds mechanical heart valves and offers diverse disease-management solutions (SJM was acquired

by Abbott Laboratories in January of 2017). Specifically, it is maintained that after months of denying the existence of a problem, SJM released patches and guidance for security vulnerabilities in its IoT medical devices. These patches address vulnerabilities in the Merlin@home Transmitter, SJM remote monitoring system of implantable pacemakers and defibrillator devices. SJM's security updates arrived concurrently with the U.S. Food and Drug Administration (FDA) and the U.S. Department of Homeland Security Industrial Control Systems Cyber Emergency Response Team separate statements detailing the vulnerabilities and advice for healthcare providers, patients and caregivers. FDA stated "The FDA has reviewed information concerning potential cybersecurity vulnerabilities associated with SJM's Merlin@home transmitter and has confirmed that these vulnerabilities, if exploited, could allow an unauthorized user, i.e., someone other than the patient's physician, to remotely access a patient's RF-enabled implanted cardiac device by altering the Merlin@home transmitter. The altered Merlin@home Transmitter could then be used to modify programming commands to the implanted device, which could result in rapid battery depletion and/or administration of inappropriate pacing or shocks" (FDA, 2017). Even though, the vulnerabilities were originally found in August 2016 by security researchers (MedSec), regarding security flaws in the encryption of the RF protocol used by the transmitter's remote monitoring system, as well as other backdoors to the device, SJM denied the vulnerabilities' existence and subsequently filed an ongoing lawsuit against security researchers for defamation through false medical device security findings. Therefore SJM has been harshly criticized for not addressing the vulnerable devices for five months since the initial disclosure, while the announced patches are not considered to address larger problems, like the existence of a universal code that could allow hackers to control the implants, as stated by Justine Bone, CEO and director at MedSec, based in Miami.

After that, and taking into account the fact that Internet-connected devices are a growing security concern FDA finalized its prolonged draft guidance on management of Cybersecurity in Medical Devices (FDA, 2016) which explicitly contains nonbinding recommendations, however includes a section on "Medical Device Cybersecurity Risk Management". Apparently, the fact that the recommendations are not legally enforceable is a basis for concern. FDA in its issuance has focused on medical device cybersecurity after devices have left the factory, recognizing and mainly proactively addressing ongoing vulnerabilities. Moreover, they have compiled a list of criteria for defining active participation by a manufacturer in an Information Sharing and Analysis Organization (ISAO), so that to bring together similar manufacturers and prompt them to share details about security risks and responses as they occur.

Having in mind the RMIAS referenced in previous section, we can maintain that FDA in favor of security development life cycle issued an earlier set of recommendations (FDA, 2014) for manufacturers in

order to build cybersecurity protections into medical devices as they're being designed and developed. "Protecting medical devices from ever-shifting cybersecurity threats requires an all-out, lifecycle approach that begins with early product development and extends throughout the product's lifespan" and "In today's world of medical devices that are connected to a hospital's network or even a patient's own internet service at home, we see significant technological advances in patient care and, at the same time, an increase in the risk of cybersecurity breaches that could affect a device's performance and functionality," wrote Schwartz (Schwartz, 2016).

Moreover, the U.S. Federal Trade Commission (FTC) also addresses the security risks associated with internet-connected devices with its recently launched prize competition that challenges the public to create a technical solution ("tool") that consumers can use to guard against security vulnerabilities in software found on the Internet of Things (IoT) devices in their homes, with the prize for the competition to be up to $25,000, ($3,000 available for each honorable mention winner(s)), and deadline for registering and submitting entries on **May 22, 2017** at https://www.ftc.gov/iot-home-inspector-challenge.

Another interesting term nowadays is Cyber Insurance. However it is not a recent trend and has its roots back in 1990s. In general, cyber insurance was created to address risk that cannot be reasonably mitigated by security measures and is in brief a product that has been created to counter residual risk associated with the information systems of asset owners. On this subject, ENISA identified that Cyber Insurance could accelerate growth for the European market, thus issued a comprehensive report (ENISA, 2016) in order to raise awareness for the most impactful market advances, by shortly presenting the most significant cyber insurance developments for the past four years (2012-2016) and to capture good practices and challenges during the early stages of cyber insurance lifecycle. The immature state of cyber insurance in Europe combined with the General Data Protection Regulation (GDPR) adopted on April 2016 (Regulation EU, 2016) and Network and Information Security (NIS) directive adopted on July 2016 (Directive EU, 2016), have created a growth anticipation embraced by enabling an informative product development and adoption.

4.3 Dejected or excited of Security Challenges?

Challenges and concerning issues are deliberately presented first in our approach, on the grounds that we would like to grab the attention of the reader regarding pervasive new technologies and deployments, whereas is our firm belief that we have just scratched the iceberg in our previous section. Broadminded people should re-consider the domain of Cyber - Security, primarily individually and

afterwards socially. It is not in the scope of our analysis to develop a personal cyber - security strategy, although individuals are encouraged to do so.

However ominous it may appears the environment of Cyber - Security in the new computationally demanding era, we should admit that the common feature of all aforementioned scientific papers and organizational handbooks is that even though the new Big Data era has bore brand new security and privacy issues, has also yielded enormous power that could be harnessed in a security enhancing manner. Advantageously manipulating Big Data is a significant challenge with numerous concerns, but it is also a brilliant chance for security analysts to gain momentum.

Finally it's up to the interest of security researchers, students and business executives to lean on the latest technological advances and deploy them in favor of our indispensable Security. Either dejected or excited, the world will probably go on and we have to turn to our internal unique mindset, as the last frontier, yet enormous weapon against unethical practices, in order to set our personal minimal to optimal contribution to the whole premise of Cyber - Security.

5 Demystifying Security in Big Data Era

Every newbie entering the information security "province" will undoubtedly face a bunch of frustrating security terminology mixed with commercial security products into a complicated notion of security. It is daunting to strive to comprehend the full concept of information security, however the RMIAS is a great model to prioritize things in your mind and distinguish security goals from countermeasures or security taxonomy. But if traditional data was examined by a traditional security point of view (like RMIAS and CIA-triad), what about the voluminous big data? Does big data equal big security? It is already proven that big data has revolutionized security like many other significant fields. From improved security development cycle to advanced security countermeasures, big data has to offer many more than widely expected. By extensively utilizing big data, new security terms have been coined and old terms have been given a newer definition. This is the proof of the extreme impact factor of these data. But before going too far, we should begin with some rudimentary notions, definitions, historical sequence and misconceptions here, conducive to forthcoming sections.

5.1 Intelligence, Business Intelligence & Big Data Analytics

At the forefront of all subsequent notions, it seems critically important to understand the fundamental difference between raw or unprocessed information and real intelligence. Information underpins Intelligence but there are significant distinctions between them. Information is comprised of raw,

unfiltered, unevaluated, incomplete, inconsistent data, collected from a variety of sources, at the same time Intelligence is processed, sorted, evaluated, complete, accurate and consistent information, collected from a variety of trusted and reliable sources. Dissemination of valuable intelligence is the essence of all vivid organizations.

However, there are numerous definitions of intelligence, even Central Intelligence Agency of USA – (CIA) has adopted its own one in 1994; "Intelligence is the collecting and processing of that information about foreign countries and their agents which is needed by a government for its foreign policy and for national security, the conduct of non-attributable activities abroad to facilitate the implementation of foreign policy, and the protection of both process and product, as well as persons and organizations concerned with these, against unauthorized disclosure" (CIA, 1995).

Clarifying the term **Intelligence** in a more Computer Science context, we should acknowledge that this term has been used by researchers in artificial intelligence since the 1950s. As mentioned in (Chen et al., 2012), **Business Intelligence – (BI)** became a popular term in the business and IT communities in the 1990s and in the late 2000s, **Business Analytics** was introduced to represent the key analytical component in BI (Davenport, 2006). BI has been used as an umbrella term to describe concepts and methods to improve business decision making by using fact-based support systems. BI includes the underlying architectures, tools, databases, applications, and methodologies (Lim et al., 2013). BI's major objectives are to enable interactive and easy access to diverse data, enable manipulation and transformation of these data, and provide business managers and analysts the ability to conduct appropriate analyses and perform actions, heavily relying on various advanced data collection, extraction, and analysis technologies (Turban et al., 2008)(Watson et al., 2011).

SearchSecurity, a branch of TechTarget online enterprise (NASDAQ: TTGT), boiled the definition of **Big data analytics** down to the process of examining large data sets to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information, while the analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits. However, the primary goal of big data analytics is to accommodate more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs (Rouse et al., 2014).

Gartner in 2014 comprehensively approached the notion of advanced analytics, stating that **Advanced Analytics** is "the analysis of all kinds of data using sophisticated quantitative methods (for

example, statistics, descriptive and predictive data mining, simulation and optimization) to produce insights that traditional approaches to business intelligence (BI) — such as query and reporting — are unlikely to discover." (Herschel et al., 2015).

5.2 Business Intelligence and Analytics (BI&A)

In the mostly cited paper about Business Intelligence and Analytics BI&A (Chen et al., 2012), the evolution of BI&A is examined thoroughly along discrete stages, starting from BI&A 1.0 in 1990s with its relational databases and analytical techniques, relying mainly on statistical methods and data mining, moving forward to BI&A 2.0 in 2000s with its HTTP-based Web 1.0 systems and Web 2.0-based social and crowd-sourcing systems, text & web analytics for unstructured web online content, fetching BI&A 3.0 to our days with mobile analytics' proliferation, wireless network sensors' analysis in blossom and promising ubiquitous Internet-enabled devices, equipped with RFID, radio tags and barcodes or QRcodes, enabling the IoT vision, creating in parallel challenging needs in analytics. Most importantly, one major application of BI&A and big data analytics is considered to be the domain of Security and Public Safety, seriously impacting Crime Analysis, Computational Criminology, Terrorism Informatics, Open-source intelligence & Cyber security, while generating unique opportunities for Criminal Association Rule Mining and Clustering, Criminal Network Analysis, Spatial-Temporal Analysis and Visualization, Multilingual Text Analytics, Sentiment and Affect Analysis, Cyber Attacks Analysis and Attribution.

Although it is a commonplace to meet Business Intelligence and Analytics (BI&A) as a unified term which treat big data analytics as a related field among it (Chen et al., 2012) (Lim et al., 2013) (Chiang et al., 2012), nowadays the need to be more precise and accurate about these buzzwords around analytics have urged analytics industries to classify and study separately each aspect of analytics. It is therefore praised every endeavor to sort out notions in easily comprehensible conceptual frameworks. To this end, **RapidMiner**, a well-known company for its open source data science platform, in one of its occasionally issued white papers (RapidMiner, 2014), presents interestingly the overarching domain of higher Analytics in a tree-based approach, branching out in two basic sub-fields, Business Intelligence – (BI) and Advanced Analytics. In general, Analytics is considered to refer to skills, technologies, applications and practices for continuous iterative exploration and investigation of data to gain insight and drive business planning, while the sub-area of Business Intelligence, focuses on using a consistent set of metrics to measure past performance, guide business planning, and by querying, reporting and OLAP (online analytical processing) can answer questions like "what happened", "how many" and "how often". Meanwhile, Advanced Analytics, goes beyond BI by using sophisticated modeling techniques to predict future events or discover patterns otherwise undetected, and can answer questions like "why", "what if", "what next" (prediction)

and “what is the best” (optimization). This white paper primarily focuses on the differences between BI and Advanced Analytics, realizing that significant discrepancy indeed does exist. The orientation, methods, knowledge generation and business initiative is briefly the case. Specifically, BI corresponds in the past or at present, uses reporting (KPIs, metrics), automated monitoring/alerting (thresholds), dashboards, scorecards, OLAP (Cubes, Slice & Dice, Drilling) & ad hoc query, generates manual knowledge and fulfills a reactive initiative. Whereas advanced Analytics corresponds in the future, uses Predictive Modeling, Data Mining, Text Mining, Multimedia Mining, Descriptive Modeling, Statistical / Quantitative Analysis, Simulation & Optimization, generates automatic knowledge and fulfills a proactive initiative.

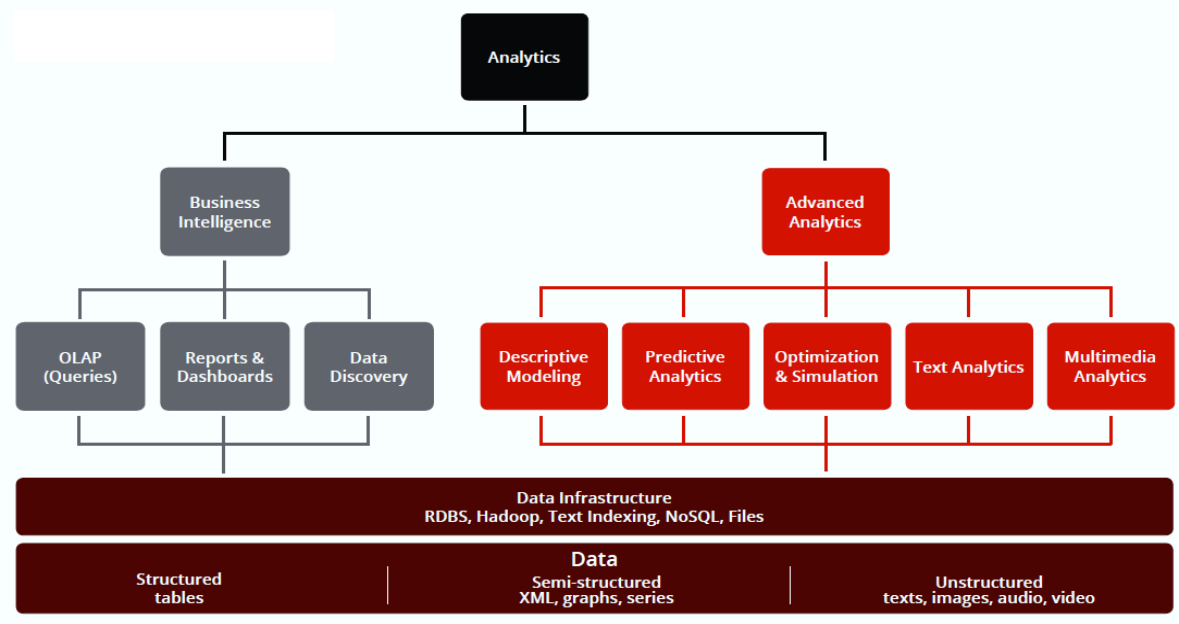


Figure 9. The overarching domain of higher Analytics (RapidMiner, 2014)

5.3 Firewalls and Intrusion Detection Systems – Intrusion Prevention Systems

Traditionally inside the world of computer and communication security, while scratching notions like secure networking and trusted or untrusted networks, firewalls and intrusion detection systems come into play.

Firewalls, either network-based or host-based, monitor and control the incoming and outgoing network traffic based on predetermined security rules. Firewalls fall under one out of three distinct categories; Packet filters, Circuit-level Gateways and Application-level Gateways. Packet filters with its access control lists (ACLs), inspect IP headers and TCP/UDP headers at a low level of OSI layered model and permit or deny traffic. Either stateless or statefull the level inside OSI model remains the same. Application-level Gateways however, provide deeper inspection upon packets and its payload, bringing more complex rules, yet more accurate control (Mavridis, 2015).

Intrusion detection systems (**IDS**) are devices or software applications that monitor a network or systems for malicious activity or policy violations. Any detected activity or violation is typically reported either to an administrator or collected centrally using a security information and event management (SIEM) system. The most common categories of IDS are network IDS (**NIDS**) analyzing for example incoming network traffic, and host-based IDS (**HIDS**) monitoring operating system files. Moreover IDS can be divided into signature-based detection, recognizing malicious patterns, and anomaly-based detection, detecting deviations from benign traffic, often deploying machine learning algorithms. Finally, some IDS can respond to detected intrusions, so these ones are typically referred to as an intrusion prevention system **IPS** (Mavridis, 2015).

Next-Generation Firewall (**NGFW**) is an integrated network platform that is a part of the third generation of firewall technology, combining a traditional firewall with other network device filtering functionalities, such as an application firewall using in-line deep packet inspection (DPI), an intrusion prevention system (IPS) and other techniques, such as TLS/SSL encrypted traffic inspection, website filtering, QoS/bandwidth management, antivirus inspection and third-party identity management integration (Pescatore et al, 2009).

5.4 Security Intelligence (SI)

At the forefront of advanced security is **Security Intelligence (SI)**. But what exactly is SI?

As IBM Corporation (NYSE) clarifies: "Security intelligence is the act of gathering and analyzing all the security-related network and file data within your organization to provide greater visibility into who is doing what with whom and which resources are accessed in the process. Similar to business intelligence, security intelligence involves the automated processing and analysis of large volumes of data. However, unlike business intelligence, the goal is not to gain a deeper understanding of a market or identify related customer preferences and buying patterns. Rather, security intelligence seeks to understand what is normal with respect to user, application and data-access behaviors occurring on your network so that when abnormal conditions arise, they can be rapidly detected and investigated". Regarding the difficulties of SI, IBM continues with: "Security data is cryptic and it comes from assets (servers), endpoints, devices, applications, users, packet captures, network scans, threat intelligence sources and now even collaboration websites. Security data is also abundant—it comes at you like a raging river and must be analyzed in real time or your team will simply drown in the details" (IBM, 2016).

TechTarget, succinctly points out that "*Security intelligence* is the information relevant to protecting an organization from external and internal threats as well as the processes, policies and tools designed to gather and analyze that information. In this context, Intelligence is actionable and provides decision-

making support and possibly a strategic advantage. In general, SI is a comprehensive approach that integrates multiple processes and practices designed to protect the organization" (Rouse, 2015).

LogRhythm, a security intelligence company, in its white paper (Petersen, 2015), boldly states that "There's no standard definition for Security Intelligence; it means different things to different companies". However they define **Security Intelligence** as "the ability to capture, correlate, visualize, and analyze forensic data in order to develop actionable insight to detect and mitigate threats that pose real harm to the organization, and to build a more proactive defense for the future. Users of Security Intelligence will shorten their Mean-Time-to-Detect and Mean-Time-to-Respond, extend the value of current security tools, and discover previously unseen threats through advanced machine analytics". LogRhythm consider that the role of Security Intelligence is to deliver actionable insight into potentially damaging threats, with supporting forensic data and contextually rich intelligence. To this end, Security teams must be able to quickly evaluate threats to determine the level of risk as well as whether an incident has occurred. Ensuring that analysts have as much information as possible to make good decisions critically enables their efficiency and decision support processes.

5.5 Log management, Security information and event management (SIEM), Network behavior anomaly detection (NBAD), Network forensics

Studying the basic components of Security Intelligence - (SI) you will definitely get stumbled upon to some of its core elements: **1) Log management**, the collective processes and policies used to administer and facilitate the generation, transmission, analysis, storage and ultimate disposal of the large volumes of log data created within an information system, **2) Security information and event management (SIEM)**, an approach to security management that seeks to provide a holistic view of an organization's information technology (IT) security, by gathering security-related events from end-user devices, servers, network equipment and specialized security equipment like firewalls, antivirus or intrusion prevention systems and by forwarding these events to a centralized management console, which performs inspections and flags anomalies, **3) Network Behavior Analysis** (**NBA**), a constant passive monitoring of network activity, including bandwidth and protocol usage, flagging unusual actions or deviations of normal benchmarked network operation. It is considered an additional measure to conventional IDS/IPS, firewalls, antivirus and anti-spyware solutions, defending ultimately a network's perimeter. NBA keep an eye on internal network, aggregating data from all possible sources to support analysis. More specifically inside NBA, **Network behavior anomaly detection (NBAD),** performs relentless monitoring of a network for unusual events or trends, with appropriate software to track critical network characteristics in real time and generate alarms if strange

events or trends are detected that could indicate the presence of threats, **4) Risk management**, the process of identifying, assessing and controlling threats to an organization's capital and earnings, including financial uncertainty, legal liabilities, strategic management errors, accidents, natural disasters and information technology (IT) security threats, and finally **5) Network forensics**, the capture, recording, and analysis of network events for the purpose of discovering the source of security attacks or other problem incidents.

Therefore, let's elaborate here more on some of these elements. We will dive into **SIEM** and there is a good reason behind that, which is going to be unfolded later on this paper, when ENISA's considerations are to be presented. Afterwards we will discuss surpass **Network Behavior Analysis**, because on a later section there is a clear understating of what is going on to that specific field, bringing into light the **UEBA** definition and future potential. Finally we will examine **Network Forensics** as a novel concept for confronting premium attacks. We chose not to elaborate on Log or Risk Management, due to the encompassing nature of other elements, which are underpinned by Log management and provide inherently a kind of Risk analysis and Management. However, there are great extensive tools about log management and risk management, however they are out of the scope of our paper.

5.5.1 Security Information and Event Management (SIEM), Security Information Management (SIM), Security Event Management (SEM)

According to (Grahn et al, 2017), the functionality of a general SIEM can be expressed by the "**5 Cs**"; **Collection** (Collect vast amount of logs from disparate sources, while data transmission to SIEM needs to be confidential, authenticated, and reliable), **Consolidation** (The desirable normalization of log data into a standard format), **Correlation** (Intensive processing of log events to unravel a nasty attack), **Communication** (three ways of interaction, either set alerts to administrators, or sent reports at a predetermined time or actively monitoring the SIEM), and **Control** (When in use data normally stored online, otherwise stored in normalized, compressed and/or encrypted form) (Karlzén, 2009). The whole premise of SIEM was built upon the necessity to examine all the data from a centralized single point, so that abnormal patterns could be immediately identified, analyzed and remediated.

Furthermore TechTarget's white paper and website (Scarfone, 2016), SIEMs basically centralize logs of security events and analyze them. The underlying principle of SIEMs is that relevant data about an enterprise's security is produced in multiple locations and being able to look at all the data from a single point of view makes it easier to spot trends and see patterns that are out of the ordinary. SIEM products first appeared in the mid-2000s, merging SIM (security information management) and SEM (security event management) functions into one security management system. A SEM system centralizes the storage and

interpretation of logs and allows near real-time analysis which enables security personnel to take quicker defensive actions, while a SIM system collects data into a central repository for trend analysis and provides automated reporting for compliance and centralized reporting. By bringing these two functions together, SIEMs provide quicker identification, analysis and recovery of security events as well as fulfilling an organization's legal compliance requirements. Generally speaking monitoring, documenting and analyzing system events is crucial component of security intelligence (SI). By utilizing SIEMs, attacks can be detected, security reconfigurations can be placed autonomously, even confirmed security breaches can be stopped. Except incident handling, SIEMs streamline compliance reporting, through robust centralized logging and customizable reporting capabilities supporting common compliance efforts, such as Health Insurance Portability and Accountability Act of 1996 (HIPAA), Payment Card Industry Data Security Standard (PCI DSS) and Sarbanes-Oxley Act (SOX).

Therefore, SIEMs are used either for security compliance or for incident detection & handling or both (an extremely difficult and complex task). SIEMs are made available mainly through four architectures, with its own pros and cons; as software installed on an on-premises server or on-premises hardware appliance or on-premises virtual appliance or as public cloud-based service. An important aspect of SIEM is the transferring of log data, from each log source to the SIEMs. Two approaches usually apply, **agent-based** (a software agent installed on each host generating logs, extracting, processing and transmitting it to the SIEM server) and **agentless** (the log-generating host directly or via intermediate logging server, like syslog server, transmits its logs to the SIEM server). Most SIEMs work by deploying multiple collection agents in a hierarchical manner to gather security-related events from end-user devices, servers, network equipment, even specialized security equipment like firewalls, antivirus or intrusion prevention systems. The collectors forward events to a centralized management console, which performs inspections and flags anomalies (to identify anomalous events, it's important that the SIEM administrator first creates a profile of the system under normal event conditions). SIEMs can be rule-based or employ a statistical correlation engine to establish relationships between event log entries. In some systems, pre-processing may happen at edge collectors, funneling only certain events to the centralized management node, reducing the information transferring and storage (the danger of filtering out relevant events too soon unfortunately remains). SIEMs offerings spans form light solutions with basic log management & reporting to advanced analysis techniques and features. They may require some hardware and/or software purchases, whereas cloud-based SIEMs are usually based on usage fees. Subsequently, a distinguishing feature of SIEMs is integration. No value can be produced unless the system can readily receive and parse log data from a wide variety of security log sources. Extensive customization could be needed (not to mention custom code to translate a source's log data into a format that the SIEMs

can understand and process). Once upon a time, SIEMs thought to be only appropriate for large organizations with advanced security needs and capabilities. At that point, security administrators and analysts in their agony to support incident detection and immediate response would duplicate network security logs in a centralized location in order to gain access to all log files through a single console and potentially correlate events across multiple log sources. Since that time, SIEMs have evolved to a core security component for every small to medium-sized organization. As the number of security log sources has gradually soared, so has the dire need to view, analyze and report on those log entries within a central single console. Not only for cementing internal security, but also for accommodating compliance needs (special legal requirements of generating reports proving the organization's accountability).

5.5.2 Network Forensics – Forensic Analytics

As Gary Palmer noted during the first Digital Forensic Research Workshop in 2001 *"The use of scientifically proven techniques to collect, fuse, identify, examine, correlate, analyze, and document digital evidence from multiple, actively processing and transmitting digital sources for the purpose of uncovering facts related to the planned intent, or measured success of unauthorized activities meant to disrupt, corrupt, and or compromise system components as well as providing information to assist in response to or recovery from these activities".*

Therefore, Network forensics is a sub-branch of digital forensics relating to the monitoring and analysis of computer network traffic for the purposes of information gathering, legal evidence, or intrusion detection. Unlike other areas of digital forensics, network investigations deal with volatile and dynamic information. Network traffic is transmitted and then lost, so network forensics is often a pro-active investigation (Casey et al, 2014).

Network forensics basically relates to Law Enforcement Agents (LEAs). Analysis of captured network traffic can include tasks such as reassembling transferred files, searching for keywords and parsing human communication such as emails or chat sessions. As (Simson, 2002) pointed out, Network Forensics Systems come in two forms; **i) "Catch-it-as-you-can"**, where all packets passing through a certain traffic point are captured, stored and subsequently analyzed in batch mode. This is the storage intensive approach, which requires large amounts of storage, **ii) "Stop, look and listen",** where each packet is analyzed in a rudimentary way in memory and only certain information is saved for future analysis. This is the computationally intensive approach, which requires a faster processor to keep up with incoming traffic. Probably, both approaches require significant storage capabilities; therefore the need for occasional erasing of old data may arise. Wireshark and tcpdump (mostly unix-like OS) are well-known programs for data

capture and analysis. Finally we should mention the "**Wireless forensics**", which is a sub-discipline of network forensics, mapping the network forensics to their wireless dimension.

A different aspect came into light form (Grahn et al, 2017), where they state; "Network forensics is a subset of both information security and big data. By capturing, recording, and analyzing network events the source of an attack can be found. Prevention of future attacks may be achieve". But what exactly comprises such kind of forensics? Anton Chuvakin, a Research Vice President in Gartner for Technical Professionals (GTP) Security and Risk Management group, outlined a broader definition of Network forensics, encompassing: Full packet capture, Data retention for a period of time, Access to captured data via search tools, Packet header analysis, Packet content analysis, including session viewing, application protocol analysis, file extraction, etc (Chuvakin, 2013). Therefore Network Forensics can be used to uncover security issues through big data analytics and automatically act upon them.

5.6 Advanced Persistent Threat (APT)

This paper should opt for a sinister attack vector in order to subsequently prove the basic assumption that big data techniques and analytics could help confront with it. Therefore among serious kinds of malicious attacks we chose APTs, one of the most sophisticated and ominous contemporary threats.

Defining APT is not an easy task. As (Edwards et al, 2015) states, many definitions have been outlined to demystify APTs, but usually they lack consistency or clarity. Thus, for testing purposes the following quote seemed quite practical; "*A targeted attack is an infection scenario executed against a limited and pre-selected set of high-value assets or physical systems with the explicit purpose of data exfiltration or damage".* In this interesting article writers highlight the paradox of anti-APT product testing, in a vague defined attack vector, whereas real working environments are infeasible without human intervention and realistic testings approach social studies rather that technical ones. Not to mention the unknown nature of these attacks that has to be tested! Lowering customer expectations is crucial when dealing with APT defenses, sometimes accepting that threat blocking may happen unknown time after the initial compromise. However impossible the idea of testing zero-day attacks may sound, there is a smart way to gain deeper knowledge. Instead of finding zero-day vulnerabilities, insert your own vulnerabilities into open source software and then build exploits for these weaknesses, by which you do not compromise general public and simultaneously you can test defender's protection ability from novel and unseen threats.

The definition of APT used by the U.S. National Institute of Standards and Technologies (NIST) is comprehensive enough; "An adversary that possesses sophisticated levels of expertise and significant resources which allow it to create opportunities to achieve its objectives by using multiple attack vectors

(e.g., cyber, physical, and deception). These objectives typically include establishing and extending footholds within the information technology infrastructure of the targeted organizations for purposes of exfiltrating (i.e., transporting it from internal networks to external drop servers) information, undermining or impeding critical aspects of a mission, program, or organization; or positioning itself to carry out these objectives in the future. The advanced persistent threat: (i) pursues its objectives repeatedly over an extended period of time; (ii) adapts to defenders' efforts to resist it; and (iii) is determined to maintain the level of interaction needed to execute its objectives" (Kissel, 2013).

Understanding the APT lifecycle is critically essential, while trying to confront them (Brewer, 2014). APTs are considered advanced on the grounds that the attacker usually writes custom zero-day malware and exploits, designed to target a specific organization, frequently launching advanced social engineering attacks. The persistent element is explained by the fact that these attacks are extremely patient and methodical till they achieve their goal. APTs are particularly renowned for lying dormant for months, waiting the opportunity to strike (such as the notorious "Stuxnet" malicious worm which attacked Iranian nuclear facilities by compromising the programmable logic controllers (PLCs), being undetected for almost two years while collecting vital information and finally caused the fast-spinning centrifuges to tear themselves apart). In the aforementioned paper, the discrete phases of the lifecycle that an APT passes through are presented, namely the Reconnaissance phase, the Compromise phase, the Maintaining phase, the Lateral Movement phase and finally the Data Exfiltration phase. The analysis of these phases extends the scope of this paper. Therefore, before attempting to stop an APT, it is crucial to determine where in the "APT lifecycle" it resides, because even though two APTs are almost always completely different, most of them usually follow a common pattern, which if fully understood can underpin the quick mitigation of the impact.

Moreover, (Grahn et al, 2017) explicitly maintain that: *"*Big data analytics tools are particularly suitable for APT detection. To detect APT attacks, collection and correlation of large quantities of diverse data including internal data sources and external shared intelligence data is a necessity" while mentioning that "Network security analytics use big data software technologies like *Hadoop* and *Apache Mahout* which extend the *MapReduce* programming model…on top of which the open-source real-time processing system *Storm* can process big data streams in a distributed, scalable and fault-tolerant way".

This is exactly the approach of our paper too. We should examine whether novel tools like Apache Storm could be utilized in the defense endeavor against malicious attacks, and we will clearly explain why the comprehension of the real – time processing of big data is more than necessary and indeed complementary to other big data software technologies.

5.7 Cyber Threat Intelligence (CTI)

Rouse et al, (2015) at TechTarget's web-based reference and self-education tool about IT, named WhatIs, which was firstly created by IBM technical writer Lowell Thing and in 1999 was acquired by Techtarget, cite that "Threat intelligence (TI), also known as Cyber Threat Intelligence (CTI) is organized, analyzed and refined information about potential or current attacks that threaten an organization. The primary purpose of threat intelligence is helping organizations understand the risks of the most common and severe external threats, such as zero-day threats, APTs and exploits. Although threat actors also include internal (or insider) and partner threats, the emphasis is on the types that are most likely to affect a particular organization's environment. TI includes in-depth information about specific threats to help an organization protect itself from the types of attacks that could do them the most damage. TI is a component of security intelligence and, like SI, includes both the information relevant to protecting an organization from external and inside threats as well as the processes, policies and tools designed to gather and analyze that information. TI services provide organizations with current information related to potential attack sources relevant to their businesses; some also offer consultation service".

It's worth noting that even though CTI sharing has become inescapably important, the currently available CTI formats (YARA, Open Indicators of Compromise - OpenIOC, Structured Threat Information eXpression - STIX), can not suffice to describe complex contemporary attack or threat patterns like in case of sophisticated attacks. The standardized formats offer wider integration of CTI into prevention and detection systems primarily due to sharing and automated processing of TI. By directly sharing such information, it is no longer necessary to wait a security system vendor to release a signature or an attack-pattern update, especially nowadays with publicly available CTI feeds in a standard format (Hail a Taxii, X-Force Exchange - IBM). But, however detailed the description of an attack can be through a standard format, most CTI feeds only share basic information (IP's, hash sums, static patterns & domains) and cannot cope with complex and sophisticated attacks in a reliable and precise manner. For instance in multi-step attack activities like seen in APTs, all partial attack steps have to be unquestionably identified, because it could be misleading to trace only separate steps which if checked independently could be considered harmless. To this end, existing CTI formats do not support full description of these complex patterns, therefore in (Ussath et al., 2016) an extension to STIX format is presented and explained on the basis that complex patterns are based on attribute relations that can be described through the tagging of relevant attributes and the specification of object attribute relations.

5.8 Cyber Threat Intelligence – (CTI) vs Cyber Threat Analytics – (CTA)

In the intense confusion of distinction between Cyber Threat Intelligence – (CTI) and Cyber Threat Analytics – (CTA) IKANOW (2015) propounds the fact that every organization, publication, or institution has its own definition of CTI. To this end, CTI is considered as data/information that has gone through some sort of evaluation process or meets preconfigured correlation rules to make it more valuable, accurate, and attributed to previously identified or new threats, being easier to incorporate into workflows, search through and analyze faster, all conducive to make manipulated information more actionable. Along this process TI feeds can be generated, being extremely valuable and offering thorough insights into vulnerabilities, exploitations, threat actors, indicators of compromise (IOCs), and much more. On the other hand CTA is considered to be the discovery, visualization and communication of meaningful patterns in data, a procedure usually executed by a cyber and risk analytics platform, gaining insights from a variety of sources, including private TI feeds, open-source data, network logs, enterprise data, social media data and more, providing the ability to easily pivot from TI into enterprise data and produce analytics to drive timely and accurate decision making. All in all, CTI and CTA platforms must work together, in order to provide a more proactive approach so that to efficiently defend against the unpredictable and constantly changing cyber threat landscape.

In the endeavor of the reader or the researcher to clarify misconceptions about quite similar security terms and related terminology, spanning from basic security goals like confidentiality and integrity to more complex ones, like specific attacks (i.e. active man in the middle (MitM) attack using cache poisoning) or countermeasures (NIDS/HIDS/NIPS/HIPS etc.), there are plenty online glossaries to opt for, however online definitions should be always taken with a pinch of salt and a paid attention to details and references. Moreover, many national or international organizational initiatives are easily found through an online search for security glossaries and a plethora of papers, articles and guidelines on security issues.

However more scientific approaches do exist and are well referenced nowadays from the Internet Engineering Task Force (IETF) and the Internet Society (ISOC), the principal technical development and standards - setting bodies for the Internet (Shirey, 2003) & (Shirey, 2007), not to mention the all time classic document “Security Architecture for the Internet Protocol” provided by the ISOC including a basic glossary in Appendix A (Seo, 2005), which really helps to fill the gaps in background vocabulary, although assuming that the reader is quite familiar with the Internet Protocol (IP), related networking technology and general information system security terms and concepts.

5.9 The dire need for immediate security stance change

If we take into account a LogRhythm's paper by Petersen (2015), IT environments have become much more vulnerable as enterprise mobility, cloud services and "bring-your-own-everything" have broken down the defensible perimeter and added layers of complexity to securing the enterprise. At the same time, the nature of cyber threats has changed dramatically. Threat actors are well organized and well funded, and many of them are known to be supported by nation states. They have sophisticated technical skills which allow these actors to create custom malware for very specific targets, and they are relentless in pursuit of their objectives. Moreover, almost anyone with a malicious intent can purchase malware and rent botnets on the Dark Web, lowering the bar for criminal entities, nation states, and terrorists to use cyber as a weapon of choice towards their intended purpose. The reality today is that for most organizations, if a motivated adversary wants to penetrate their network, they will get in.

Many organizations still continue to focus their attention on identifying and blocking threats at the defense perimeter. Unfortunately, it is evident that the traditional security tools organizations have long relied on to protect their networks (anti-virus, anti-malware, IPSs/IDSs, firewalls, application gateways, endpoint protection and so on) cannot keep up with the rapid escalation of sophisticated threats and prevention-centric strategies are constantly failing in front of intensive attacks, especially now that attackers are conducting reconnaissance to find weaknesses to strike and attacks are made stealthy to pass all preventive measures. While traditional solutions still provide businesses a basic level of protection, in today's fast-changing world, they must be used in conjunction with systems that provide steep Security Intelligence. Only with this kind of increased defense, an attack can be identified and remediated far earlier, particularly when the vast majority of APTs involve the use of zero-day malware – which point security tools often miss.

The 2015 Global State of Information Security Survey shows that the compound annual growth rate (CAGR) of detected security incidents has increased 66 % year-over-year since 2009. Survey respondents acknowledge detecting a total number of 42.8 million security incidents in 2014—an increase of 48 percent over incidents detected in 2013. That's the equivalent of 117,339 incoming attacks per day, every day, and that's only what has been detected and reported.1 One cyber security company recently estimated that as many as 71 percent of compromises go undetected (Petersen, 2015).

The way to bring visibility to the most important threats while clearing the fog of noise is with Security Intelligence (SI). Just as Business Intelligence has helped numerous organizations clear the fog of too many points of seemingly extraneous business data to find previously unknown business opportunities, Security Intelligence does much the same thing with threat information, enabling companies to clearly see the threats that matter.

Utilizing Big data for security reasons has many aspects and could yield many different use cases, like (Lan, et al., 2013) which provide ideas for network security monitor system on Big Data environment, while analyzing a large number of heterogeneous types of security incidents, focusing on the correlation algorithms and knowledge representation on Big Data environment. In that paper, Big Data were organized through data cleaning, data integration and data reduction and then proceed to the core data, using algorithms on core data, mainly fuzzy constraint correlation algorithm based on prerequisites and consequences of security events, traffic rules base on wavelets and inter-related rules correlation base on sequence pattern. A prototype system was built for typical events needing further consideration.

Striving to keep up with the evolution of security intelligence (Cates, 2015) in an era where SaaS applications account for over 50% of IT application spending, governments and enterprises are forecasted to double their spending in cloud resources by 2018, mobile usage of data is expanding and an avalanche of data is starting to flow from a myriad of connected devices realizing the Internet of things (IoT), inevitably an urgent need is generated for tools that can intelligently and proactively process the enormous volume of data and identify threats in real-time before organizations are compromised. Criminal gangs harvesting the cheap computation power and resources have created new methods of attacks that can't be detected by traditional security solutions (zero-day exploits). Relying only on firewalls, IDS/IPS or simple antivirus has become a danger. Therefore, Security intelligence tools, like SIEMs that can detect hidden usage and access patterns otherwise unavailable are necessary to neutralize threats resulted from increased hacking techniques and malicious insider compromises.

Likewise in (Bhatt et al., 2014), writers point out the operational and technical challenges that traditional SIEMs are facing, while admitting that advances in computer science will significantly impact SIEM systems and unravel potential capabilities, especially when implementing parallel and distributed computing methods and big data analytics, like a distributed correlation engine that could handle complex rules and identify actionable security information from very large event datasets.

An integral part of enterprise computer security incident response teams, a security operations center (SOC) monitors security incidents in real time. Security incident and event management systems play a critical role in SOCs collecting, normalizing, storing, and correlating events to identify malicious activities. Although SIEM systems provide a solid technical foundation for SOCs, further advances are needed to create a more adaptive, context-aware, flexible, holistic, and social solution. Fortunately, advances in storage systems (nonvolatile memories) and advances in parallel and distributed computing (big data analysis) will provide the platform for scalable analysis, like a distributed correlation engine that could handle complex rules and identify actionable security information from very large event datasets. Having all the network and host logs at the Security Analysts - SAs' fingertips is attractive because the

more information a SOC has, the better its situational awareness. However, this comes at the cost of transforming an SIEM system's data management to big data management, which turns storage, search, sharing, transfer, analysis, and visualization into challenges. One aspect of this problem is the system's inability to efficiently execute complex queries, severely limiting SAs' ability to write complex correlation rules. Most teams focus on short-term alerting functionality and ignore the long-term retention feature. SOCs usually monitor a rolling and narrow time window of events (typically an hour or two). This limits their ability to detect stealthy, slow-advancing attacks, especially APTs (Bhatt, S. et al, 2014).

Also, Virvilis et al., (2014) stated that today's major cyber threats are targeting governmental, military and industrial communication and information systems in a far more well organized way being heavily funded, sometimes falling under a state umbrella with abundant money, time and expertise. After analyzing large-scale attacks to critical national infrastructures either in USA in 2006 & 2008, Estonia in 2007 or UK in 2007, and attacks to well-known corporations like Google; Gmail in 2010, RSA in 2011 or Comodo in 2011, common features could categorize some of them as Advanced Persistent Threats (APTs). Even though intrusion detection problems are considered inherently difficult, signature-based detection approaches have been tried with unfortunately poor results. Alternative methods like behavioral or statistical approaches and machine learning techniques had discouraging results in real-time environments. Recently cloud-based approaches allowing for central collection of information and data analysis from a large user base have enabled quicker responses but limited benefits for APTs. Unique characteristics of APTs like zero-day exploits, attackers' significant time spending, perpetrators' state support and the high selectiveness of naive victims have rendered current cyber solutions ineffective or inadequate. Network/Host – based Detection Systems (NIDS/HIDS) either by signature-based or anomaly-based detection strategy fall apart when confronting APTs, because either new custom malware have not signatures yet or anomaly detection suffers from obfuscating false positives. Not to mention the legitimate resembling traffic used by malwares like SSL /TLS. Moreover, the limited time window (usually seconds) is not conducive to detection of stealthy and long malicious activity. Neither are Antivirus Products, as long as attackers are adapting their malicious code to antivirus. Full Packet Capture by inspecting captured traffic may could help, but their limited analysis and lack of integration with other systems is discouraging. Also Security Incident and Event Management (SIEM) systems present shortcomings like limited time window and centrally performed correlation. Finally, the lack of efficient integration solutions, different proprietary rules and configuration languages with individual knowledge banks and lack of open standards creates a miserable mix to defend advanced threats. Here comes big data analytics, whereas a full packet capture approach with deep packet inspection and big data analytics would allow for pointed insights and advanced correlations. The time needed for APTs should be exploited backwards by security analysts who

can harness the power of massive event correlation across huge timescales and multiple sources. Novelties of big data analytics should pave the way for more efficient previously unattainable holistic approach without time of computation constraints.

To seal the difficulties in intrusion detection systems, we can remember Cohen (1987) who introduced computer viruses and in his paper concluded that "the goals of sharing in a general purpose multilevel security system may be in such direct opposition to the goals of viral security as to make their reconciliation and coexistence impossible".

In (Fetjah et al., 2016) writers recognized that traditional intrusion detection system (IDS), intrusion prevention system (IPS), and Security Information Event Management (SIEM) can't cope with the sheer amount of nowadays generated big data and aren't capable of confronting sophisticated attacks such as Advanced Persistent Threats (APT). On that grounds, a novel solution (named Advanced Persistent Security Insights System - APSIS) was developed for Security Intelligence, focusing on an approach that relies on advantages of the traditional SIEM systems; which incorporate data aggregation, correlations, alerting, dashboards, compliance, retention and forensic analysis and then expose it to big data with the application of security intelligence in order to have more accurate view on what is happing on the infrastructure.

We can see specific benefits from big data tools and it worth referring to a case study presented by Zions Bancorporation (Darkreading, 2012). At a RSA Conference in 2012 they showed how Hadoop and BI analytics can power better security intelligence than established traditional tools. Its study found that the massive amount of data and events that they had to analyze were too much for traditional SIEMs (20 to 60min were needed to search among a month's load of data). In the new Hadoop system running queries with Hive, the same results were yielded in approximately one minute. From 20-60 minutes to 1 minute was groundbreaking result. Also, the security data warehouse driving this implementation let users mine meaningful security information from a variety of security devices, website traffic, business processes and other transactions. This incorporation of unstructured data and multiple disparate datasets into a single analysis framework is one of big data's promising features. Big data tools are also particularly suited to become fundamental for APT detection and forensics generally. To detect these attacks, we need to collect and correlate large quantities of diverse data (including internal data sources and external shared intelligence data) and perform long-term historical correlation to incorporate a posteriori information of an attack in the network's history.

There are numerous ways in which big data analytics could help achieve greater Security. An example could be the identification of all assets deployed to an organization. Asset identification helps organizations to identify and to respond quickly to any security breaches. In (Arora, et al., 2016), machine

learning based techniques are used to identify assets, based on their connectivity and results show that a viable solution can be provided in automating asset classification, especially for large datasets. When automating assets, various IoT devices can be easily maintained and controlled, mitigating the risk of device breakages, as well as allowing quickly identification and response to any security breaches or event failures.

6 Methodology

Our whole Master Thesis was basically Big Data and Security oriented. Our goal throughout this research was primarily to outline contemporary cyber security notions, which combined with the evidently ongoing Big Data era and its corresponding tools, could unfold an enormous potential. Our purpose was ultimately to raise public awareness about these issues, and finally to contribute a grain of knowledge, utilizing a novel processing paradigm for enhanced security.

Therefore our methodology was quite straight-forward. Firstly, capture the interest of our readers with big data numbers. Interestingly enough, figures depicting the vast amount of currently generated data our outstanding. Secondly, scrutinize security challenges and significant – sensitive security issues. Thirdly, clarify interchangeably misused security notions, introducing the dire need for a security stance change. Subsequently, bring our reader to the "last month's" advancements and pave the way to future deployments, taking into account thoughts of well-versed stakeholders regarding research and strategic analysis.

Afterwards, to underpin our belief for an open interconnected society, we enumerate some free and open source solutions, which encompass or complement big data stream analytics. Dive into open source projects, we undoubtedly opt for Hadoop ecosystem, and among participants we chose specifically Apache Storm.

Storm was chosen to remediate real time processing concerns at high rates in a reliable and easily deployable manner. Seeking relevant comparative references, we stumbled upon some papers benchmarking stream processing tools, and we finally picked Storm. After that, we outlined the exact features of Storm and its core functions, in order to understand our subsequent practical implementation.

Later on, we declared our practical goal, to accomplish a generally security enhanced solution utilizing a stream processing tool and cyber security intelligence feeds. To this end we were convinced that every endeavor to analyze stream data at high rates for security reasons would be helpful to prove our basic idea.

Among numerous implementations, we chose Apache Web Server, knowing that Apache was always a great solution regarding Web Servers and maintained a praise-worthy reputation about its solid functionality, with a big market share among web server software providers.

Analyzing fast generated log files of an Apache Web Server, would be beneficial to our research. Analyzing basically meant to us, some form of processing to distinguish malicious from benign users. So, we decided to analyze each and every record yielded from a web server, by parsing the whole record, extract the source IP and later, after creating our malicious – blacklisted dataset, to check if the incoming

(source) IP was listed in our malicious dataset, in order to drop every packet from that user – potential attacker to our infrastructure. So, simultaneously we created our unique and tailor-made dataset of blacklisted IPs. A clear deployment of well-known APIs for security intelligence feeds and cyber threats pulses, oriented towards our personal needs and/or our security goals.

The whole premise of our application was to maintain simplicity and clarity, enhance our security perimeter by deploying a stream processing paradigm with automatically updated security intelligence feeds, customize our approach towards our needs and provide functionality in a scalable and reliable manner.

With a dataset of 50 million records and thousand blacklisted IPs, we run our experimentation on a Vsphere Virtual Machine with 32GB RAM and 24 cores with Debian Jessie OS. The exact implementation and the interesting results can be found at the end of chapter 12.

Our application was an example application from a non exhaustive list of potential applications. It is up to the ambitious reader and/or researcher of this thesis to extend the functionalities or the application areas.

Finally, considering our results, we explain some basic assumptions and we declare our known limitations upon our approach. There we outline the current situation on relevant issues and provide a quick refutation to shed some extra light to our approach.

7 Big data for enhanced Cyber Security Intelligence

Contemporary hardware and software solutions strive to cope with big data revolution. Inherently difficult to capture, store, filter, share, analyze and visualize on current technologies, one could easily realize that big data is becoming progressively an important issue in academia, business and governments. Despite difficulties, harnessing big data could generate revenue, executive efficiency, strategic decisions, better services, as well as aid to identify trends and develop new products, all of which is covered in the data science (Chen et al., 2014). Data science studies parallel and distributed processing, similarity search, graph analysis, clustering, stream processing, search ranking, association analysis, dimensionality reduction and machine learning algorithms (Miloslavskaya et al., 2014). In this context we should incorporate security, advancing traditional data environments.

The promising differences between big data and traditional data include the data collection and aggregation procedure, the distinguishing analysis (big data analysis), the infrastructure used to store and house big data and the technologies applied to analyze, process, manipulate, visualize, report and transfer structured, semi-structured and unstructured big data from heterogeneous sources and environments. It is very much anticipated that these differences could render big data the critical factor which can unleash an unrealized potential for a variety of scientific, industrial and governmental domains. In this paper, we are primarily interested in confronting advanced security challenges via big data implementation capabilities.

First CSA (Big Data Analytics for Security Intelligence), anticipated that SIEMs apart from managing alerts from different intrusion detection sensors and rules, would aggregate and filter alarms from many sources and present actionable information to security analysts. Moreover, CSA foresaw the generation of the second generation SIEMs, truly incorporating Big Data analytics into security, with Big Data tools that have the potential to provide a significant advance in actionable security intelligence by reducing the time for correlating, consolidating, and contextualizing diverse security event information, and also for correlating long-term historical data for forensic purposes.

Later, according to well-versed ENISA and its guide about Security of big data (Naydenov et al, 2015), the most promising domains of big data applications have been already identified, among which primarily stands out **advanced security information and event management (SIEM)** and less notably **Cyber Security analytics.** In our Big Data era, second generation SIEMs can capture unstructured data relevant to enterprise security from wide variety of sources and carry out complex queries yielding results

in a timely fashion. Unstructured data was difficult to capture by utilizing first generation SIEMs. One major advantage of Big Data & NoSQL, is that provide storage in a scalable format and allow better insights while creating complex queries. Big Data offers precious added value from the correlation between unstructured data and SIEM scalability.

Mattern et al., (2014) articulate that a transformed approach to cyber security shouldn't rely solely on responding to known threats; it must also track the capabilities, intentions, and activities of potential adversaries and competitors, as they evolve, in the cyber realm. That set of information and associated functions is referred to as Cyber Intelligence. Cyber Intelligence seeks to not only understand network operations and activities, but also who is doing them, why, and what might be next. Intelligence functions for cyber security include collecting and analyzing information that produces timely reporting, with context and relevance to a supported decision maker.

Award-winning technology journalist Lemos reports (2015) that interest in advancing the analytics capabilities of SIEMs is on the rise. In our big data era, incorporating special data analysis into SIEMs is a partially met challenge. While SIEMs have promised visibility into network events and potential threats, their ability to identify true threats remains uneven. Automating the collection and management of log files has resulted in a disturbing large number of **false positives**. The high number of false positives and the continued failure to detect signs of advanced attacks remain a major problem for all security teams. Ponemon Institute (2015) outlined the scope of the false positives problem, demonstrating that among 17,000 malware alerts that large organizations face every week, only 3,218 (19%) are truly reliable and only 705 (4%) are finally investigated, due to lack of resources or in-house expertise to detect or block serious malware. Lemos wittingly remarked David Bianco's words; "You need analytics when humans cannot read the logs, which, nowadays, is just about always", highlighting the key benefits of analytics against manual human analysis and configuration. David Bianco is a security architect for security-intelligence firm Sqrrl Data, Inc. It is evident that nowadays the addition of analytics to traditional SIEMs mainly focuses on prioritizing alerts based on increasing chances that they represent threats. However, an EMA research found that 90% of organizations were able to reduce their false positives using security analytics (Lemos, 2015). Therefore, easily comes in mind the combination of advanced big data analytics with SIEMs, in order to help security analysts handle broader datasets, incorporate better analytics into their current deployments and improve threat detection as well as time management for security teams. Data analytics also offer an interactively search through security and business data sets for evidence of compromise, while revealed patterns can be automated for future detection of threats. Today, enterprises like Splunk Inc., RSA, The Security Division of EMC Corp. and Blue Coat Systems Inc. offer analytics-

focused products, while others such as AlienVault Inc. and LogRhythm Inc. have incorporated analytics capabilities.

With such analytics, security analysts and SIEM administrators can increase the quality of their alerts in several ways: corroborate events and alerts, incorporate lessons from the hunt or gain context via global threat intelligence. Using analytics to match events and alerts based on a common user, machine, process or file can help the truly suspicious incidents bubble up to the top of the analyst's priority list. An administrator logged into a variety of systems, an application contacting a server in China, or a process eating up CPU time may all be valid alerts, but an admin user logged into multiple machines all contacting a server in China and having high utilization should be at the top of a security analyst's priority list.

Furthermore trying to excel SIEMs' incident detection capabilities and to achieve significant accuracy, threat intelligence (TI) feeds are considered a much needed necessity (Scarfone, 2016). Incorporating up-to-date information on threat indicators observed around the world can make a difference in a SIEMs. A threat intelligence feed provides information on the latest detected threats, such as the IP addresses of devices being used to attack others and a SIEMs can use information from TI feeds to significantly improve its real-time analytics by making attack detection faster and more accurate and by giving the SIEM platform a stronger basis for prioritizing its actions. Some SIEM platforms use TI feeds that the vendor provides; others use third-party feeds. Undoubtedly, the quality of the information in the feed itself is certainly important, including how often it is updated, how comprehensive it is and how accurate it is. But it's also important how the SIEMs uses the TI feed, meaning that it should be just one of many considered factors for real-time analytics. An unbalanced approach may significantly increase false positives or false negatives, potentially making real-time analytics less effective. However, this is something analyzed thoroughly in the next section along with security analytics.

The growing complexity and sophistication of today's cyber threats, coupled with an ever-increasing volume of data in which key threat indicators are hidden, necessitates a more coordinated and efficient approach to threat detection and incident response. Information security teams are limited in their ability to prioritize investigations, efficiently gather evidence, centrally track progress, and quickly foster collaboration with and escalate to more qualified staff. "Fostering collaboration among multiple team members to expedite the evaluation, prioritization and response to threats has never been more important given today's complex threat landscape," said Michael Ables, senior network systems analyst at Tarleton State University (Boulder, 2015). "Security teams are struggling with alarm fatigue, too often chasing down the wrong alarms, missing the important ones, and doing all of it inefficiently" said Chris Petersen, co-founder & CTO at LogRhythm. "These latest innovations speak to LogRhythm's focus on solving the

most pressing challenge CISOs face today – quickly detecting and responding to those threats that could bring harm" (LogRhythm, 2015).

Another recognition that Big Data Analytics are indeed conducive to Cyber Security Intelligence comes from Jamal Elmellas in (Elmellas et al., 2016), whereas although referring to Threat Intelligence and its yet unrealized potential, it is acknowledged that big data analytics deployment for a variety of dynamic sources, such as social media, blogs, news feeds and especially deep up to dark web, could yield predictive intelligence capable of determining the probability of a serious cyber attack. This shift in security thinking is on the way, moving from post-incident remediation to pre-incident insights and predictions. An intelligence-led security revolution accompanied by big data does worth noting. Nowadays, it is strategically necessary to continuously collect, recognize and analyze the motivations, objectives, intentions, skills and capabilities of potential attackers likely to launch malicious acts upon the interested organization. These kinds of reconnaissance actions can adequately inform decision-makers and executives for a properly designed cyber-security budget allocation. Therefore, it becomes evident that the capstone of all security intelligence gathering and analyzing is the prescriptive aspect of cyber security intelligence, with novel recommending features and well-scheduled automated acting. Predictive and prescriptive big data security analytics seem unavoidable in the foreseeable future.

Finally, in December 2016 the business colossus IBM has issued a Security white paper, called: "IT executive guide to Security Intelligence", mentioning the undergoing transitioning from log management and SIEM to state-of-the-art Security. Specifically, "The concept of Security Intelligence is partially realized in SIEM tools, which correlate and analyze aggregated and normalized log data. SIEM is very strong from an event-management perspective and plays a particularly important role in threat detection. Comprehensive security intelligence, however, depends upon continuous monitoring of all relevant data sources across the IT infrastructure and evaluating it in context. This includes, but is not limited to, security and network device logs and flows, vulnerabilities, risks, configuration data, network traffic telemetry, packet captures, application events and activities, user identities, assets, geo-location and application content. To be fully comprehensive, there is also need for solutions that can consume log and service data from cloud applications to provide security visibility across on-premises, hybrid and cloud infrastructures". Therefore IBM proposes a sophisticated proprietary analytics engine called IBM Qradar, a Security Intelligence Platform which delivers Security Intelligence using advanced Sense Analytics. Use cases of Sense Analytics deployment include Advanced Threat Detection, User behavior analysis and insider threat monitoring, Risk and vulnerability management, Compliance reporting, Securing the cloud, Forensics investigation and Incident Response (IBM, 2016).

Big Data combined with Cloud Computing provide many possibilities, but also poses limitations and threats. As long as mobile users tend to transfer data storage and processing outside their device, security and privacy becomes more challenging. Examples do exist to maintain that beneficial integration of big data and cloud can be achieved as noted in

8 Big data for enhanced Cyber Security Analytics and Platforms

The widespread adoption of IoT solutions, big data analytics, mobile and cloud computing, has expanded the security perimeter of all modern organizations. These novel technologies have produced a massive amount of threat data to be monitored, examined and analyzed. Moreover, to turn all this data into meaningful intelligence, new tools are needed, beyond traditional SIEMs, in order to integrate and analyze disparate types of enormous in volume data, structured, semi-structured and unstructured.

In addition to that, a more fascinating motive can be found among security visionaries. That is moving from post-incident remediation and mitigation acts to pre-incident insights, predictions and preparation, which is considered a focal point for cyber security. This kind of shift along a timeline of a malicious attack is widely desired among businesses and academia, with predictive and prescriptive analytics take precedence over all traditional security approaches. To this end, Security Analytics are gaining momentum again.

In a recent scientific paper a Taxonomy of Security analytics was proposed (Grahn et al 2017), wherein after acknowledging the Role of Big Data in Network Security they propose their Analytics Taxonomy, which include *descriptive analytics (or data mining)* for identification of network security threats, *diagnostic analytics* for forensics, *visual analytics* valuable in the abstraction of multivariate temporal data, *predictive analytics* for proactive intrusion prevention, *prescriptive analytics* for the best action using optimization, simulation and heuristics. Subsequently, applying their proposal to network security analytics another taxonomy regarding security analytics was identified: *Descriptive security analytics (*network activity logs, SIEMs, and anomalies in network behavior - closely related to security intelligence), *Diagnostic security analytics (*identifies security attacks, detects anomalies, and carries out forensic investigations), *Visual security analytics (*raw packets and data records, security events in IPS/IDS, firewalls, VPNs, and anti-malware software, output from vulnerability scanners, network events in switches, routers, servers, and network hosts, network applications' logs), *Predictive security analytics (*estimates targets, sources, and methods of potential attacks or events), *Prescriptive security analytics*

*(*compare and optimize solutions, like protection against malware, firewall configuration, protection of network communication, recovery after security incidents).

Gartner (Ahlm et al, 2016), in its subsequent figure (Figure 8) shows the four different types of analysis that are commonly performed, each of which can yield benefits to a wide range of disparate roles within the security operations function of detecting, investigating and responding to a breach.

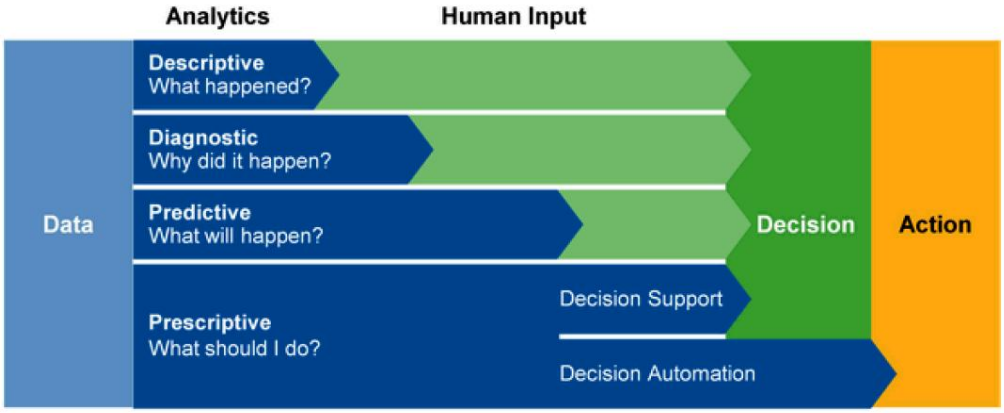


Figure 10. The Four Common Data Analysis Functions (Ahlm et al, 2016)

Regarding the notion of Cyber Security Analytics we should primarily consider that Big Data is changing the security landscape (Naydenov et al, 2015); new approaches arise as various capabilities are enabled, like analysis of long-term historical trends and predictive analysis. By collecting and analyzing historical data on a large scale, becomes possible to identify and probe an attack, its initial date and time, the exact steps of the attacker and many more. Even if they did not detect the original attack, they can go back to carry out a historical correlation in the database to identify and examine the attack. Meaningful intelligence can be generated through efficient manipulation of collected data. Complicate queries can be executed immediately, receiving quick valuable results. New threat services and analytics tools are introducing automated approaches and high-performance security monitoring systems that keep up with complex real-time analytics to improve detection rates and overall performance. New era of utilizing Big Data for predictive analytics is already present. Proactively dealing with threats is a significant feature of big data analytics.

In (Cardenas et al, 2013) & (CSA, 2016) details are presented on how and why the security analytics landscape is changing with the introduction and widespread use of new tools leveraging large quantities of structured and unstructured data (big data) to improve information security and situational awareness. The deployment of Big Data analytics is praised, especially in analyzing financial transactions, log files and network traffic to identify anomalies and suspicious activities and to correlate multiple sources of information into a coherent view. However, here we should clarify that security analytics it's not

in its infancy. It has been around for many years. Data-driven information security dates back to bank fraud detection and anomaly-based IDSs. Although analyzing logs, network packets and system events has been a problem seeking a consistent resolution for decades, conventional technologies couldn't support long-term and large-scale analytics. In the past retaining large quantities of data wasn't economically feasible, therefore event logs and other computer entries were deleted after a fixed period (usually 60 days), while performing analytics and complex queries on large, unstructured datasets with incomplete and noisy features was inefficient. Even popular SIEMs couldn't analyze and manage unstructured data. On the contrary, nowadays big data applications are becoming part of security management and play a pivotal role in cleaning, preparing and querying heterogeneous, incomplete or noisy data efficiently. Needless to mention here, that the management of large data warehouses has always been an expensive and computationally intensive function. However, Hadoop framework and other big data tools are now enabling the deployment of large-scale, reliable clusters and therefore new opportunities are emerging to process and analyze immense data. Likewise large credit card issuers have conducted large-scale fraud detection for decades; however, the custom infrastructure to mine big data for fraud detection wasn't beneficial till recent advancements. All in all, the major advantage of big data is that it can facilitate an affordable security infrastructure. Technologies, like Hadoop ecosystem (Pig, Hive, Mahout and RHadoop), stream mining, complex-event processing and NoSQL databases, are main enablers of large-scale analysis on heterogeneous datasets at unprecedented speed. These novel technologies are transforming security analytics in various ways, either by collecting data massively from many internal and external enterprise sources (such as vulnerability databases) or by performing deeper analytics on the data; or by providing a consolidated view of security-related information; or finally by achieving real-time analysis of streaming data. However, it remains important that Big Data tools will continue to require system architects and analysts to have a deep knowledge of their system in order to properly configure the Big Data analysis tools.

An exponential increase in the type, frequency and complexity of cyber attacks was brought by the rapid growth of the Internet (Industrial, Semantic, Social or just Internet of Things). Traditional Cyber-security solutions already in place to counteract these attacks are gradually rendered obsolete while Big Data implementations over computer networks are becoming a commonplace. To this end, Researchers in (Mahmood, et al., 2013) explain how corporate research is now focusing on Security Analytics, i.e., the application of Big Data Analytics techniques to cyber-security. Analytics can assist network managers particularly in the monitoring and surveillance of real-time network streams and real-time detection of both malicious and suspicious (outlying) patterns. Such behavior is envisioned to encompass and enhance all traditional security techniques. In the aforementioned paper, state-of-the-art cyber-security issues are

highlighted in the new era of big data, as well as incapability of current security solutions to encompass real-time big data network streams. Moreover, Security Analytics with Big Data Analytics are praised to provide a "richer" Cybersecurity context by separating "normal" from "abnormal", distinguishing legitimate user generated patterns from suspicious or malicious ones. Insights gained when deploying security analytics are achieved with the aid of a diversified variety of sources, classified though into two main categories. Passive sources (Computer-based data, Mobile-based data, Physical data, Human Resource data, Travel data, SIEM data, Data from external sources) and relating to real-time or Active sources (Credential data, One-time passwords, Digital Certificates, Knowledge-based questions, Biometric identification data, Social media data). Finally it is proved that Security Analytics (the application of Big Data Analytics techniques) can derive actionable intelligence and insights from streams in real-time, which is rapidly becoming a strong need for cyber-security setups.

Forrester Research states that monitoring applications, databases, endpoints and network devices creates enormous log volumes that traditional SIMs have struggled to manage. The advent of big data and advanced analytical techniques have ushered in a new era for security monitoring, giving birth to SA platforms. "SA platforms use big data technology and machine learning to rapidly examine events, looking for anomalous activity that could be indicative of a breach, active malware, or other malicious activity". Forrester specifically defines an SA platform as: "A platform built on big data infrastructure to converge logging, correlating, and reporting feeds from security information management (SIM), security solutions, network flow data, external threat intelligence, and diverse endpoints and applications. The SA platform uses this information and machine learning techniques to provide real-time monitoring and facilitate the rapid incident detection, analysis, and response". Traditional SIMs strain under the volume and velocity of big security data (logs, network metadata, flow data, events, and alerts) and by using structured databases and ingesting specific data feeds are extremely difficult utilized. As a consequence, SA platforms (either on-premises or SaaS-based) have prevailed by residing on a big data engine that can aggregate and search through many disparate data types, able to ingest unstructured data and allow analysts to pivot on any/all data, making them especially valuable for threat hunting and investigations. Not to mention the high number of alerts that SIMs produced and the extraordinary to copy with number of false positive alerts that reduced the actual visibility. Eventually, SA helps to overcome the false positive problem and provide more meaningful alerts to analysts, while the ability of SA platforms to observe behaviors throughout the network provides a more complete and accurate picture of users - devices interaction. Technologies such as network analysis and visibility (NAV) and user behavior analytics (UBA) or security user behavior analytics (SUBA) provide input for SA tools that give security teams visibility to quickly understand relationships and activities in the network (Blankenship, 2016). Finally Forrester, acknowledging the dire

need for faster threat detection and remediation and mainly predictive analytics, predicted that SA platform adoption will accelerate in the next two years, especially in five categories: in highly regulated industries (subject to PCI), already breached companies, B2B firms competing on Intellectual Property, large B2C companies regularly cyber-attacked and government agencies handling sensitive data.

Eric Ahlm, research director at Gartner says "Breach detection is top of mind for security buyers and the field of security technologies claiming to find breaches or detect advanced attacks is at an all time noise level," (Ahlm, 2017) "Security analytics platforms endeavor to bring situational awareness to security events by gathering and analyzing a broader set of data, such that the events that pose the greatest harm to an organization are found and prioritized with greater accuracy."

8.1 User Behavior Analytics (UBA)

Moreover, Gartner (2015) states that User Behavior Analytics (**UBA**) is another example of security analytics that is already gaining great attention. UBA systems allow user activity to be analyzed (like a fraud detection system would monitor user's credit cards), and are effective at detecting meaningful security events, such as a compromised user account and rogue insiders. UBA systems can analyze massive amounts of data, not only user profiles (like devices, geolocations etc.) but also many more data points, providing enhanced analytics and increasing the accuracy of detecting a breach.

UBA was initially defined by Gartner in 2014 (Litan et al., 2014) as a cyber security process about detection of insider threats, targeted attacks and financial frauds. So, UBA encompassed patterns of human behavior and application of algorithms and statistical analysis to detect meaningful anomalies indicative of potential threats, tracking users, rather than devices or security events. To this end Big data technologies (i.e. Apache Hadoop etc) have played a pivotal role, increasing UBA functionality by allowing analysis of petabytes in order to detect insider threats and APTs.

“Security systems provide so much information that it's tough to uncover information that truly indicates a potential for real attack. Analytics tools help make sense of the vast amount of data that SIEM, IDS/IPS, system logs, and other tools gather. UBA tools use a specialized type of security analytics that focuses on the behavior of systems and the people using them. UBA technology first evolved in the field of marketing, to help companies understand and predict consumer-buying patterns. But as it turns out, UBA can be extraordinarily useful in the security context too” (Johnson, 2015).

8.2 User Entity Behavior Analytics - (UEBA)

In September 2015, Gartner following the market’s substantial growth in 2015, evolved the category from UBA to User Entity Behavior Analytics - (UEBA) (Bussa et al, 2016), turning basically UBA into UEBA which includes devices, applications, servers, data, or anything with an IP address.

UEBA has five main technology components; Data Analytics (advanced analytics, rules, machine learning, deep learning), Data Integration (structured, semi-structured, unstructured), Data Presentation/Visualization (link analysis, time series & trend analysis, queries & reports across users - entities), Source Systems and Applications Analyzed (on-premises or in the cloud) and finally Service Delivery Method (on-premises or cloud-based). UEBA has enhanced Security by incorporating profiling and anomaly detection based on machine learning. Using advanced analytics, UEBA pinpoint threats and improve signal-to-noise ratio across multiple monitoring systems or information feeds used into platforms, whereas can keep pace with the increasing volume and complexity of security events. Relevant vendors use packaged analytics to monitor users' and entities' activity (user movements, access to assets and the context of that access) in order to form behavioral groups based upon common activities, correlate behaviors, detect anomalies and finally discover security infractions and malicious or abusive users. The addition of letter "E" in the term UEBA recognizes the fact that other entities besides users are often profiled in order to more accurately pinpoint threats, in part by correlating the behavior of these other entities with user behavior (below figure).

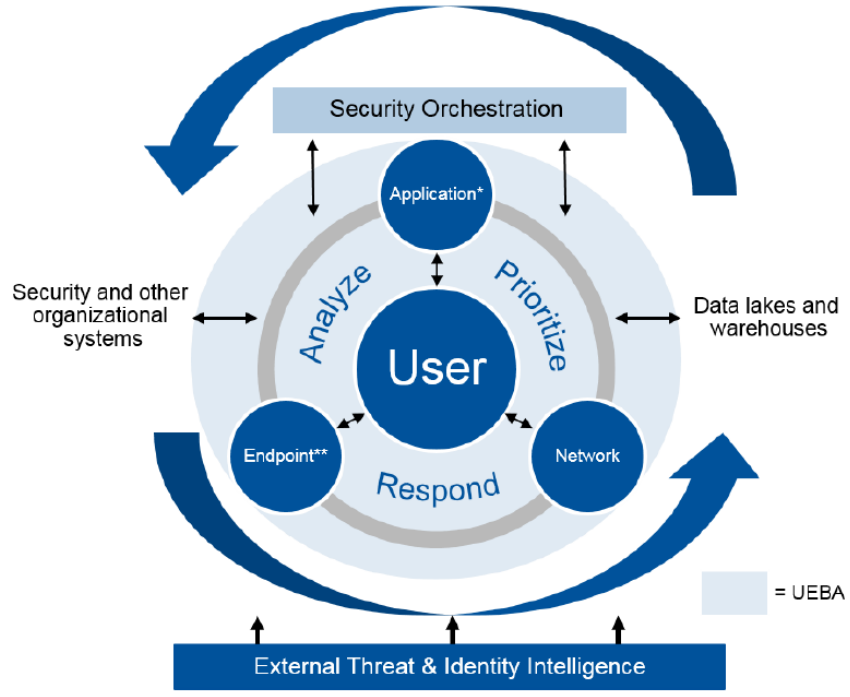


Figure 11. UEBA defined (Bussa et al., 2016)

8.2.1 UEBA market and strategic insights

In the same report Gartner maintains that despite almost $80 billion spent globally on security, attackers are still getting through organizational defenses, on the grounds that even though alerts and alarms do go off in various monitoring systems, they are unfortunately ignored since they are buried among tens or hundreds of thousands of alerts. At the same time, most enterprises spend a majority of their security budget on prevention measures (firewalls, strong user authentication, intrusion prevention, antivirus systems etc), while cunning hackers have figured out how to beat these prevention systems. Additionally, attackers are often not detected while intruding on a network, because many monitoring systems generate overwhelming amount of false alarms and as a consequence intrusion alerts go unnoticed. However, recent breaches involved hackers taking over existing user accounts, malicious activities that UEBA systems are designed to detect. Detection can be more successfully accomplished using advanced behavioral analytics rather than just plain rules. Moreover, it is already rumored that in the next few years, machine learning will start migrating into deep learning, where the models learn on their own from "training data" and select which attributes and variables to key their analytics off of. Deep learning promises to disrupt the UEBA market and other sectors that rely on machine learning and advanced analytics.

After all, UEBA can primarily analyze attacks and infractions to find malicious actors, prioritize alerts that need to act on while improving alert management by correlating and consolidating alerts and respond timely by streamlining alert and incident investigations therefore reducing the time and number of staff required to investigate those alerts. However, Gartner identified that UEBA vendors primarily align with one or more of five domains; Broad-Scope Security Management (rapidly detect and analyze bad activities, improve signal-to-noise ratio, consolidate & reduce alert volume, prioritize remaining alerts, facilitate efficient response & investigation, integration with SIEMs), Data Exfiltration (detect exfiltration, enhance existing DLP systems with anomaly detection & advanced analytics, improve signal-to-noise ratio, consolidate DLP alert volume, prioritize remaining alerts, integrate with network traffic & endpoint data), Identity Access Management (monitor & analyze user behavior against access rights, including privileged users and service accounts), Insider Threats (monitor staff for unusual or abusive behavior, find insiders engaged in malicious activities, ingest & analyze unstructured information - email content, performance reviews, social media information), and finally Saas Security (ensure security and visibility into enterprise use of SaaS applications).

It is already evident that UEBA vendors have gained greater attention by wider customer base, while market consolidated and interest in UEBA and security analytics entrenched. Thus Gartner moved beyond its fraud-oriented focus, stating that "UEBA successfully detects malicious and abusive activity

that otherwise goes unnoticed (by existing security monitoring systems, such as SIEM and Data Loss Prevention DLP), and effectively consolidates and prioritizes security alerts sent from other systems". However, the debate among advanced SIEMs and UEBA remains, with advanced SIEMs' users to claim that sufficient visibility and low alert volume is achievable as long as SIEM rules are well tuned.

"Although SIEMs can facilitate investigations through gathering and analyzing broad security data sets for centralized detection and event prioritization, it often falls short on finding advanced advisories or attacks that don't generate a log artifact or it overall tends to be human-resource-intensive" states a newer Gartner report [Market Trends: User and Entity Behavior Analytics Expand Their Market Reach]. This fact combined with the dire need for constant security monitoring on breaches has given rise to a number of smaller, fast-growth markets that use analytic technologies to help organizations enhance their efficiency in detection and augment the investigative capabilities of SIEMs. Gartner feels that spending patterns will continue in order to extend monitoring for breach, while in (Tirosh et al., 2016), Gartner predicts that by 2020, 60% of enterprise information security budgets will be allocated for rapid detection and response approaches, up from less than 20% in 2015.

In the same report it is widespread that a consolidation of behavioral-based detection systems, such as user and entity behavior analytics (UEBA), endpoint detection and response (EDR), and network traffic analysis (NTA) is already on rails. While mature advanced analytics models become an important differentiating factor among relevant technologies. Security customers demand swift breach detection as well as quick and efficient response, consequently a market collision is emerging between behavioral-based detection systems, security orchestration and response systems.

According to Gartner UEBA market comprises only vendors that support security use cases with packaged analytics. The UEBA market does not include vendors that do not profile users and do not detect anomalies upon their behavior, excluding vendors that solely analyze endpoints and/or network behavior. Neither does UEBA includes vendors that support security use cases through data mining, user-driven data exploration and visualization, but don't provide packaged user behavior security analytics, at least partly. On the contrary, near-real-time monitoring is a definitely required capability for these products.

The UEBA market grew faster and matured more quickly than Gartner anticipated a year ago. With market revenue projected to climb to $200 million by the end of 2017, market consolidation has already begun. In July 2015, Splunk acquired UEBA startup Caspida, which had just a handful of customers, for $190 million, in September 2015, Microsoft acquired Adallom, a cloud security broker with UEBA functionality and about 100 customers, for a reported $250 million and in April 2015, HP announced Securonix technology repackaged into its HP ArcSight User Behavior Analytics product.

As Gartner acutely apprehends, providers of security monitoring technologies and services need to develop or acquire statistical analysis and machine learning capabilities to incorporate into their security monitoring platforms or services. Traditional rule-based detection technology alone is unable to keep pace with the increasingly complex demands of threat and breach detection.

Gartner finally estimates that by 2017, some UEBA technology firms will be acquired by SIEM, DLP or other large security vendors, at least 60% of major cloud access security brokers and 25% of major SIEM and DLP vendors will opt for advanced analytics and UEBA functionality, and Deep Learning will be incorporated into at least one UEBA product, with a projected increase in near future. Moreover, currently non-UEBA vendors (like RedOwl Analytics, Prelert, Cloudera, Platfora and Sqrrl) that offer big data integration, visualization and advanced analytics platforms for enhanced Security purposes are likely to convert into packaging analytics partly focusing on user behavior, whereas vendors offering network and/or endpoint-centric advanced analytics will probably add user behavior analytics soon.

Mr. Ahlm also underscores that "However, the applications or other problems that can be addressed for other security markets are still emerging and on the whole, the security industry is rather immature in the application of analytics" (Rivera et al, 2015). As security analytics platforms mature, a driving factor for their accuracy is integration. Nowadays, information about hosts, networks, users and external actors is the most common data brought into an analysis; however, the amount of context that can be brought into an analysis is truly boundless and presents an opportunity for owners of interesting data and security providers to increase their effectiveness.

Analytics systems tend to do better analyzing, in almost real-time speed and produce valuable findings. One major challenge yet to accommodate is the time window relevant security events occur. There may be an early indicator, followed hours later by a minor event, which in turn is followed days or months later by a data leakage event. These three things must be evaluated as a single incident that spans, for instance three months and the overall priority of this incident, comprised of lesser events must be higher. On these grounds, historical-data retention is a key component for advanced analytics systems. "Ultimately, how actual human users interface with the outputs of large data analytics will greatly determine if the technology is adopted or deemed to produce useful information in a reasonable amount of time," said Mr. Ahlm. "Like other disciplines that have leveraged large data analytics to discover new things or produce new outputs, visualization of that data will greatly affect adoption of the technology" (Rivera et al., 2015).

In a recent press-release (HPE, 2017), Hewlett Packard Enterprise – HPE announced it has acquired Niara Inc., a California-based leader in the emerging User and Entity Behavior Analytics (UEBA) security market segment. Niara Inc. via Shashanka et al (Shashanka et al, 2017) has issued a scientific

paper explaining its fundamental UEBA module, providing a clear example of UEBA in enterprise use. According to naira use case, they chose to track and monitor behaviors of users, IP addresses and devices, trying to automatically detect anomalous behavior, using machine learning algorithms based on Singular Values Decomposition (SVD). Therefore they created two benchmark baselines, firstly a historical baseline, relevant to a user's behavior against his own behavior over time, and secondly, a peer baseline, according to the behavior of all peers of the user. Going forward they identify a set of seven features, and by extending Mahalanobis Distance (Mahalanobis, 1936), claim that can automatically recognize malicious activities, such as command and control, internal reconnaissance, lateral spread, privilege escalation and exfiltration. Also, events and alerts were appealingly scored by two values, a severity score and a confidence score. Combining both values an event can be escalated to alert. In conclusion, the followed approach was basically focused on server-access behavior anomaly detection based on SVD.

Another enterprise UEBA example is from Fortscale. Fortscale boasts about its award-winning UBA solutions which combine expertise from the Israeli Defense Force's elite security unit, advanced machine learning, and big data analytics to provide rapid detection and response to malicious user behaviors that truly matter, automatically and dynamically analyzing real-time and historic user behavior to identify and prioritize the highest-risk user access and activities associated with applications, devices, and services on network. However, Fortscale interestingly enumerated the reasons to adopt and utilize Advanced UBA (Fortscale, 2016); As briefly pointed out UBA can **(1)** Dramatically reduce data breaches & cyber attacks, giving security teams an ever-increasing degree of context and accuracy to enable simpler, smarter security operations and the insight and agility to easily and rapidly identify threats, accelerate investigations and neutralize security threats, **(2)** Automatically discover attackers and rogue employees, using advanced machine learning which automatically analyze per-user and peer-group behaviors across dynamic timeframes and statistical categories to rapidly pinpoint anomalies and instantly detect insider attacks, compromised credentials, and suspicious access to sensitive data, **(3)** Assess prioritized alerts & investigate in minutes, limiting potential threats and making security analysts more effective, by quickly focusing on the most important tasks via scoring risks and delivering prioritized basic alerts to any syslog interface, **(4)** Leverage & optimize your existing security infrastructure, linking disconnected data silos, through tight integration with SIEMs, analytics engines and Hadoop-based architecture, enable analysis of hundreds of millions access events across various applications, turning a voluminous amount of historic and real-time log data into actionable intelligence.

Key findings by Gartner are that (i) Advanced analytics are being integrated into security markets after rule- and signature-based prevention systems struggled to detect or stop most security breaches over the past few years, (ii) Common use cases for advanced security analytics include detecting cyber threats,

insider threats, data exfiltration and monitoring high-privilege user accounts for abuse or misuse, (iii) Most firms in traditional security markets (like SIEMs, endpoint protection and identity and access management – IAM), are challenged by new players offering advanced analytics in their domains. Taking these into account, Gartner recommends focusing on specific security monitoring and threat detection use cases, such as insider threat or access analytics, to find the most appropriate usage for security analytics and seek vendors with advanced analytics embedded in packaged offerings where initial tuning efforts can be reduced to a few weeks in order to show detection results. After all Gartner is already "seeing vendors with future roadmaps that combine deep learning, supervised learning and anomaly detection techniques to deliver the most effective analytics" (Bussa et al., 2016).

In a comprehensive e-publication of SearchSecurity.com, another approach of Advanced Security Analytics is met, wherein is maintained that this kind of analytics hold the promise of protecting enterprise organizations against the most sophisticated attacks, including APTs (Ashford et al., 2016). SearchSecurity.com believes that the best way to understand advanced security analytics is to consider security technologies in the pyramid of Maslow's hierarchy of needs. At the base are **protection systems**: firewalls, secure web gateways, antiviruses/antimalware, identity and access management, and data loss prevention. Tools at this level scan for malicious code and monitor for unauthorized access to various resources, protecting a specific type of resource, system or attack vector. Tools essential, but not sufficient (coordination and mining is needed). At the next level up we find **detection and monitoring systems**, such as SIEMs and IDS/IPS. These tools take a holistic view of the enterprise, monitor and manage information across multiple resources and systems, including security tools and they integrate into protection systems. All these tools collect and filter enormous amounts of data, but analyzing it isn't humanly possible and that's where advanced security analytics comes in. **Advanced Security Analytics** is a higher layer of systems that integrate into existing products and automate the analysis via machine learning and big data techniques. This approach enables information security professionals to take action on the issues that truly represent a breach or threat. There are several different types of products that fall within the category of advanced security analytics. Firstly, **Security operational intelligence** tools, which permit users to uncover connections between events, secondly **Behavioral threat analytics** (BTA) tools that analyze user's behavior, devices and systems in the environment to uncover anomalous behavior that may represent a threat. Security Analytics tools typically get their inputs from the data and feeds of other tools and systems, including SIEMs, IDS/IPS and a range of others (firewalls, secure web gateways and the like). Other tools focus on contextual analysis of insider threats and other on the translation of threat to business risk. In general though, analytics tools attempt to tease out incidents, or patterns of incidents that represent an actual threat, distinguishing between harmless and malicious anomalous behavior. Through machine

learning, artificial intelligence and integration with other security solutions, they can reduce false positives by several orders of magnitude (from 500 or more a day to two or three real threats) so that security professionals can address manually. But they won't stop here. The next frontier of advanced security analytics is **predictive analytics** not just detecting threats as they occur, but accurately predicting the threats that will hit tomorrow and the business risk they will pose. That will, in turn, enable security professionals to move from reactive to proactive mode.

8.3 Advanced Security Market Predictions

Lastly, within a comprehensive report by Gartner (Ahlm et al., 2016), the consolidation of behavioral-based detection systems is predicted, such as user and entity behavior analytics (UEBA), endpoint detection and response (EDR), and network traffic analysis (NTA). Moreover the acceleration of Advanced Analytics Capabilities into the Security Market is laid down, proving the severe impact of Advanced Analytics on Security. Outlining the advanced analytics maturity model, reveals the focus of robust players on solving a wider range of business challenges through analytics. This model has three legs. Firstly, data which spans from structured, internal, siloed data to unstructured, external till hybrid, integrated. Secondly decisions from Ad hoc, batch, offline analytics to pervasive, real-time, embedded analytics. Lastly there is analytics, moving from descriptive, to diagnostic, later predictive and finally prescriptive analytics. However, variations in application, scope and capacities set the boundaries for maturity, therefore some vendors focus on ingesting new or complex data types, others focus more on decision support, and others still focus on the analytic process itself. Using the advanced analytics maturity model becomes an important tool for differentiating various security behavioral detection technologies. "Although each leg is important, what can be done (in terms of analysis) to the data so that a better outcome (decision or action) can be made is often the focus of technology discussions regarding analytics in security" says Gartner.

Research and advisory firm Forrester Research (Nasdaq: FORR) considers that **Security Analytics (SA)** has garnered a lot of attention during the past few years. However, marketing hype and misunderstandings regarding SA have confused the market, making it difficult for security and risk (S&R) leaders to make informed decisions. "It's all about visibility. The primary purpose of SA is to provide centralized visibility across the environment for quick threat detection and resolution". Even though S&R professionals initially deployed security information management systems (SIMs) to satisfy compliance requirements for security monitoring and log management, nowadays struggle to use SIM for threat detection and visibility; "SIMs will evolve into SA platforms and will help S&R professionals detect unknown threats while monitoring behavior inside the network, not just at the perimeter" (Blankenship,

2016). Cautiously, Forrester warns of the fallacy on Security Analytics. For the past years, many marketers resonate recklessly the buzzword of SA, claiming SA solutions, even though they do not provide any SA component among their proposition. Thus, be aware of misconceptions, distinguishing standalone security analytics components from an integrated and comprehensive SA platform, which must incorporate many features for the whole security environment.

Characteristic	Description
Speed of transaction analysis	The ability to analyze a threat event and return a decision about it in real time — leading to automated control enforcement or escalation for event response
Amount of data analyzed	Petabytes. Big data tools make it possible to store log data for longer periods of time (even forever). Security analysts can use this historical data for investigations, threat hunting, and trending.
Big data infrastructure	Diversity and volume of data necessitate a big data infrastructure.
Event correlation process	Context-based, adaptive, and risk-based threat detection
Integrated platform	The platform must include the ability to capture network analysis and visibility (NAV), threat intelligence, security user behavior analytics (SUBA), and SIM data, and then present that information in a centralized console.
Machine learning	Supervised and unsupervised machine learning methods detect anomalous behavior without the need for prewritten rules.
Risk computation models	Statistical-based and rules-based risk and security event modeling
Entity and link analytics	Entity and link analytics — evaluating network, host, and endpoint devices link together
Statistical probability methods	Probabilistic models to determine the likelihood of a breach
Workflows	Built-in workflows for analysis and investigation
Visualization	Tools to graphically display behaviors and relationships between devices

Figure 12. Security Analytics key Characteristics by Forrester (Blankenship, 2016)

8.4 Strategic Cyber Security Intelligence

We chose to close this section with a reference to the notion of strategic cyber security intelligence. Unfortunately, our thesis lacked the time and the scope to dive deeper into these complicate issues. However, we managed to emerge a compelling scientific paper (Borum et al, 2015) which highlights that many cyber security discussions and warnings emphasize on tactical cyber intelligence supporting the "on-the-network" fight. Nonetheless, at the same moment, strategic and operational levels of cyber intelligence receive less attention (Dog et al., 2016). A lack of emphasis on the strategic aspect of Cyber Intelligence reveals the current weakness in providing strategic intelligence. This paper undoubtedly shed light to the importance and role of strategic cyber intelligence to support risk-informed decision-making, ultimately

leading to improved objectives, policies, architectures and investments to advance a nation or organization's interests in the cyber domain.

9 Open Source Solutions for enhanced Security - Big data Batch and Stream Analytics

It is evident that the security business landscape is chaotic. Many entrenched players do exist with a variety of advanced cyber security propositions and loyal clients; whereas new players like tech start-ups keep entering this market place taking gradually small bites. The reader is encouraged to seek the competitive forces inside that security market. However, the strategic analysis of enterprise solutions left outside the scope of our paper. We focused on our belief that Open Source solutions are becoming viable alternatives, and alluring free editions of commercial solutions do exist, even though precaution should be taken in such scenarios so that to avoid a possible enterprise "lock-in'.

Here a clarification is necessary, regarding advanced Security software solutions, which realize big data stream analytics and provide real – time analytics for Security reasons. Apparently, when contemplating Security in a broader sense, every platform which underpins stream processing could be potential be used for enhanced Security. Complex algorithms, can be combined and applied upon marketing, HR (human resources) or sales figures similarly to security logs, events or captured packets. Therefore, many well-known companies offering big data stream analytics platforms can be utilized for the sake of some aspect of security. As Forrester, the research and advisory firm, declares: "Nonsecurity analytics vendors are entering the security space. Vendors like SAS and FICO, better known for their analytics prowess than for their security chops, have entered the market. These may be especially attractive in the financial services space where these vendors already have traction in fraud prevention" (Blankenship, 2016).

Lastly, yet significantly, we should demystify novel enterprise costly Security solutions and declare that these propositions internally implement open source frameworks in order to construct and provide their security platforms and/or applications. Therefore knowledge of the below mentioned frameworks, are considered a must have qualification to firms that need to harness big data, and we feel ambivalent for the existence of firms do not need or care for big data insights.

9.1 Apache HADOOP ecosystem

Till now, the reader should have already understand from scientific references the revolutionary nature of Hadoop, but here we should give a brief description and make a clear distinction between core hadoop and its wide ecosystem, in order to go beyond base hadoop modules and familiarize later ourselves with Real –Time analytics in conjunction with Hadoop.

Apache Hadoop is an open-source software framework used for distributed storage and processing of dataset of big data using the MapReduce programming model. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures. (Apache org)
The genesis of Hadoop was the "Google File System" paper, published in October 2003 (Ghemawat et al, 2003) and later another one from Google – "MapReduce: Simplified Data Processing on Large Clusters" (Dean et al, 2008). Hadoop has its origins in Apache Nutch project (today an extensible and scalable open source web crawler software project), itself a part of the Lucene project. The Apache Nutch initially included Hadoop, but in 2006 were separated when moved to the new Hadoop subproject (implementing MapRedue and Dfs – distributed file system). The initial code factored out of Nutch project consisted of about 5,000 lines of code for HDFS and about 6,000 lines of code for MapReduce. The first committer to add to the Hadoop project was Owen O'Malley (in March 2006); First version of Hadoop was released in April 2006 and till now it grows and evolves through contributions being made to the project.

In his definitive and comprehensive guide about Hadoop, White gave much more details on the ecosystem (White, 2012). Regarding the project he mentioned that it was created by Doug Cutting, the creator of Apache Lucene, the widely used text search library. As for the origin of the Name "Hadoop", he points to the project's creator, Doug Cutting, explaining that he named the project after his son's toy, a stuffed yellow elephant. Doug remarked: "Short, relatively easy to spell and pronounce, meaningless, and not used elsewhere: those are my naming criteria. Kids are good at generating such". Projects in the Hadoop ecosystem also tend to have names that are unrelated to their function, often with an elephant or other animal theme like "Pig" (White, 2012).

The core of Apache Hadoop consists of a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a MapReduce programming model. Hadoop splits files into large blocks and distributes them across nodes in a cluster. It then transfers packaged code into nodes to process the data in parallel. This approach takes advantage of data locality, where nodes manipulate the data they have access to, allowing the dataset to be processed faster and more efficiently. We should also mention that the Hadoop framework itself is mostly written in the Java programming language, with some native code in C and command line utilities written as shell scripts, even though today any programming language can be used with "Hadoop Streaming" to implement the "map" and "reduce" parts of a user's program.

The official Apache website (Apache Hadoop, 2017) distinguishes core hadoop components from other related components on top of Hadoop:

Core Hadoop Components:

- **Hadoop Common** → The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS)** → A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**→ A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce** → A YARN-based system for parallel processing of large data sets.

Other Hadoop - related projects at Apache include:

- **Apache Ambari** → A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
- **Apache Avro** → A is a remote procedure call and data serialization system. It provides a serialization format for persistent data and a wire format for communication between Hadoop nodes and from client programs to Hadoop services.
- **Apache Cassandra** → A scalable multi-master database with no single points of failure. It is linearly scalable, low latency, highly available and fault tolerant without compromising performance or mission-critical data.
- **Apache Chukwa** → A data collection system for monitoring and managing large distributed systems. Built on top of Hadoop Distributed File System (HDFS) and Map/Reduce framework it inherits Hadoop's scalability and robustness. Apache Chukwa also includes a flexible and powerful toolkit for displaying, monitoring and analyzing results to make the best use of the collected data.
- **Apache HBase** → A scalable, distributed database that supports structured data storage for large tables.
- **Apache Hive** → A data warehouse infrastructure that provides data summarization and ad hoc querying.
- **Apache Mahout** → A Scalable machine learning and data mining library. Apache Mahout provides three major features: A simple and extensible programming environment and framework for building scalable algorithms, A wide variety of premade algorithms for Scala + Apache Spark, H2O, Apache Flink, Samsara, a vector math experimentation environment with R-like syntax which works at scale

- **Apache Pig** → A high-level data-flow language and execution framework for parallel computation.
- **Apache Storm** → **A** fast, scalable and fault-tolerant distributed realtime computation system, reliable to process unbounded streams of data, doing for realtime processing what Hadoop did for batch processing. Used with any programming language, has many use cases: realtime analytics, online machine learning, continuous computation, distributed RPC, ETL, and more.
- **Apache Spark** → A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- **Tez** → A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases. Tez is being adopted by Hive, Pig and other frameworks in the Hadoop ecosystem, and also by other commercial software (e.g. ETL tools), to replace Hadoop MapReduce as the underlying execution engine.
- **Apache ZooKeeper** → A high-performance and reliable coordination service for distributed applications. A centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.
- **Apache Phoenix** → An open source, massively parallel, relational database engine supporting OLTP for Hadoop using Apache HBase as its backing store.
- **Cloudera Impala** → Cloudera's open source massively parallel processing (MPP) SQL query engine for data stored in a computer cluster running Apache Hadoop.
- **Apache Flume** → A distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of log data.
- **Apache Sqoop** → A command-line interface application for transferring data between relational databases and Hadoop
- **Apache Oozie** → A server-based workflow scheduling system to manage Hadoop jobs.
- **Kafka** → A distributed streaming platform

When defining **Hadoop** Ecosystem, a holistic picture assists our understanding about the whole concept and relevant dependencies. Shubham Sinha, a Big Data and Hadoop expert working as a Research Analyst at Edureka, an interactive e-learning platform, offering multiple online courses, provides such a picture.

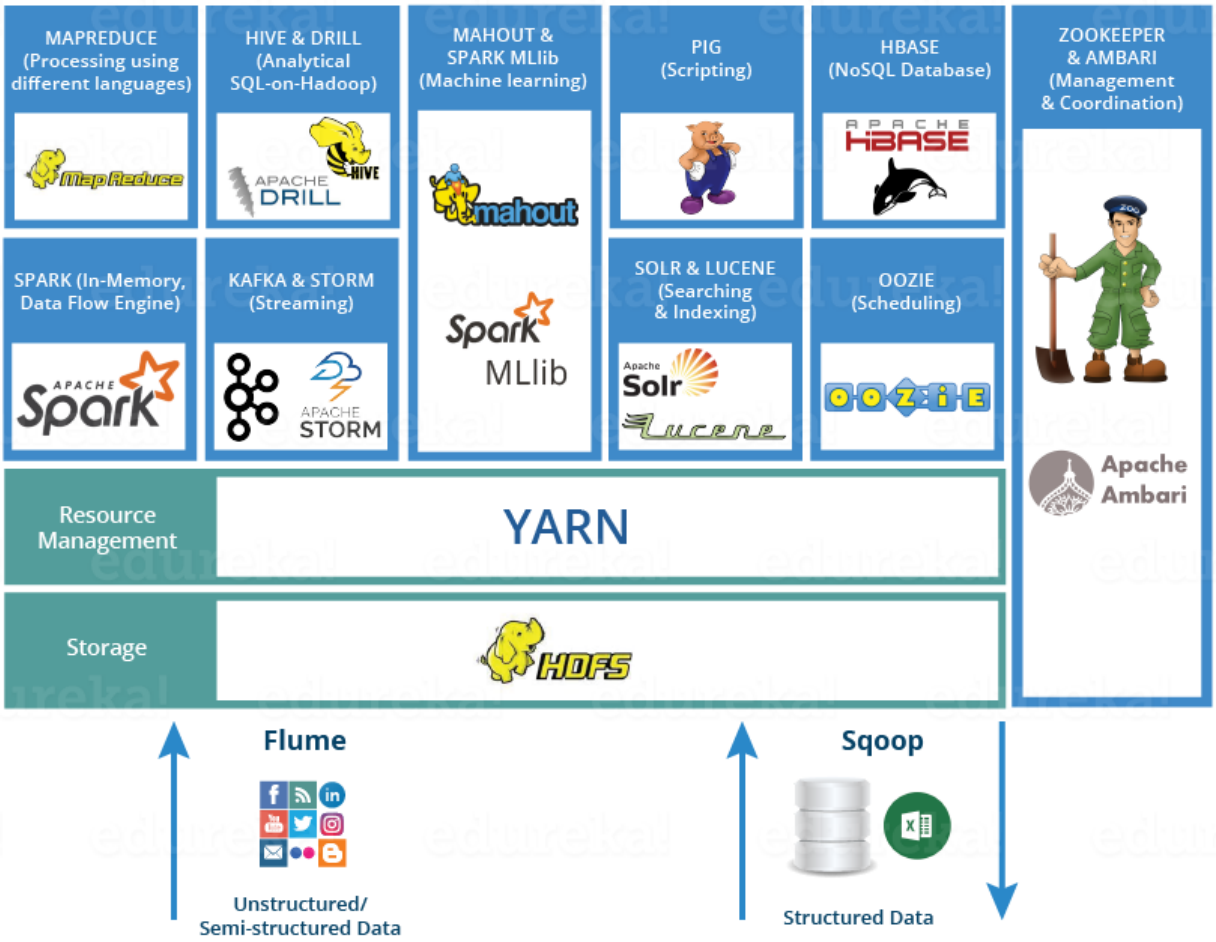


Figure 13. Hadoop Ecosystem (Shubham, 2016)

Shubham briefly enumerate some possible Hadoop components to Edureka blogs:

HDFS -> *Hadoop Distributed File System (core Hadoop)*

YARN -> *Yet Another Resource Negotiator(core Hadoop)*

MapReduce -> *Data processing using programming (core Hadoop)*

Spark -> In-memory Data Processing

PIG, HIVE-> *Data Processing Services using Query (SQL-like)*

HBase -> *NoSQL Database*

Mahout, Spark MLlib -> *Machine Learning*

Apache Drill -> *SQL on Hadoop*

Zookeeper -> *Managing Cluster*

Oozie -> *Job Scheduling*

Flume, Sqoop -> *Data Ingesting Services*

Solr & Lucene -> *Searching & Indexing*

Ambari -> *Provision, Monitor and Maintain cluster*

All in all, nowadays the term **Hadoop** has become a synonym of the whole ecosystem, in which many more additional software packages can be installed on top of or alongside core Hadoop depending on our needs and use cases. So, the reader / researcher is free to deploy its own implementation of Hadoop.

9.2 Stream Analytics

9.2.1 Commercial big data stream analytics

Then we can merely enumerate some commercial big data stream analytics for security: IBM, Software AG, Azure Stream Analytics, DataTorrent, StreamAnalytix, SQLstream Blaze, SAP Event Stream Processor, Oracle Stream Analytics, TIBCO's Event Analytics, Striim, Informatica, WSO2 Complex Event Processor, SAS Event Stream Processing, Cisco Connected Streaming Analytics. The reader in encouraged to seek more information about stream analytics and their potential.

9.2.2 Open source big data stream analytics

Open source big data stream analytics alternatives, include Apache Flink, Apache Spark Streaming, Apache Samaz and Apache Storm. Flink combines distributed stream processing with batch data processing and is considered a fault tolerant dataflow engine, providing several APIs for creating applications which use Flink. Spark let its users to construct scalable and fault tolerant streaming applications like the way they write batch jobs. Samza except being a stream processing framework, encompass resource management capabilities by using Apache Kafka messaging system and Hadoop YARN. Storm is a distributed real-time computation system for unbounded stream of data, resembling Hadoop batch capabilities but for stream data. Details on Storm are elaborated later.

9.3 All in one Real-time big data security Projects

Fortunately, Open Source software that could analyze threat data in excessive amounts while incorporating all latest techniques can be found. Here we highlight a comprehensive ongoing Security Open Project which immediately captured our interest.

- ***From OpenSoc to Apache Metron***

In 2014 while realizing that contemporary Security requires the application of big data analytics, Cisco Systems announced the OpenSOC (Open Security Operations Center). The OpenSOC project is a collaborative open source development project dedicated to providing an extensible and scalable advanced security analytics tool. It has strong foundations in the Apache Hadoop Framework and values collaboration for high-quality community-based open source development. OpenSOC is a Big Data

security analytics framework designed to consume and monitor network traffic and machine exhaust data of a data center. OpenSOC is extensible and is designed to work at a massive scale [http://opensoc.github.io/]. Basically OpenSOC is an anomaly detection and incident forensics platform, which integrates elements of the Hadoop ecosystem, including Storm, Kafka, and Elastic search, for full-packet capture, indexing, storage, data enrichment, stream and batch processing, along with real-time search and telemetry aggregation.

The developing timeline of OpenSoc was turbulent, on the grounds that after first release of OpenSoc beta in September 2014, Cisco delivered an OpenSoc Community edition in June 2015 and stopped its supporting to OpenSoc. Soon after that, Apache Metron was accepted into Apache Incubation and the first release of Metron was a reality in April 2016. Apache Metron claims to be a cyber security application framework that provides organizations the ability to ingest, process and store diverse security data feeds at scale in order to detect cyber anomalies and enable organizations to rapidly respond to them. The figure below, depicts the 4 key capabilities of Apache Metron.

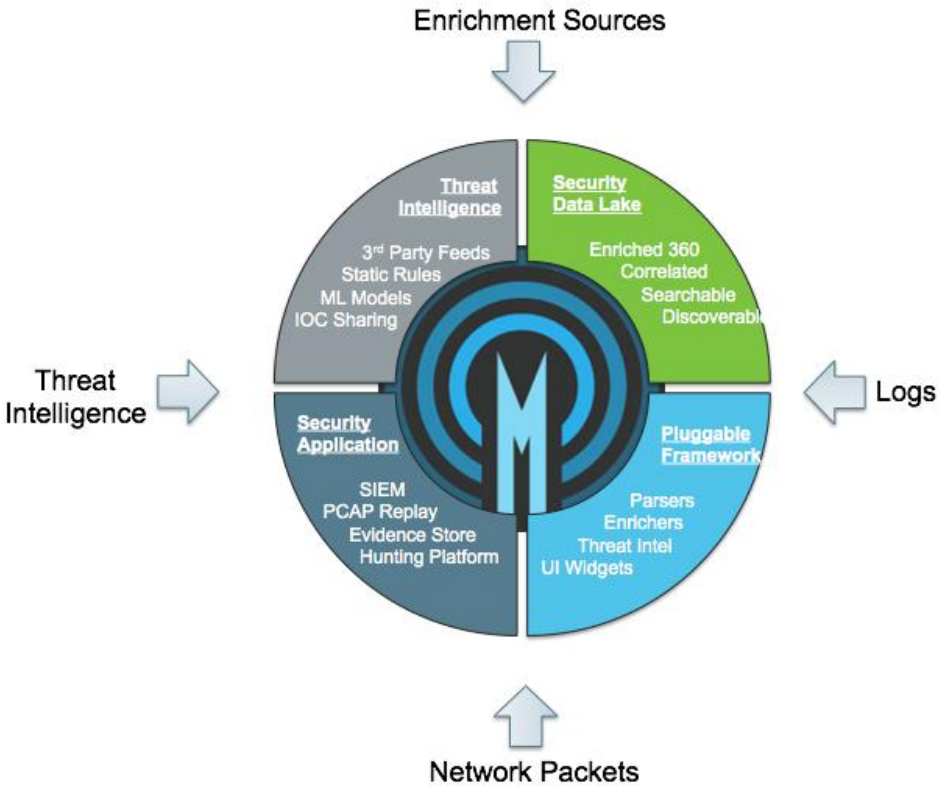


Figure 14. Metron's Core Functional Capabilities (Apache Metron, 2017)

- **Security Data Lake / Vault** – The platform provides cost effective way to store enriched telemetry data for long periods of time. Data lake enables feature engineering that powers discovery analytics and operational analytics.

- **Pluggable Framework** – Metron's platform provides a rich set of parsers for common security data sources (pcap, netflow, bro, snort, fireye, sourcefire) and permits to add new custom parsers for new data sources, new enrichment services to raw streaming data, pluggable extensions for threat intelligence feeds, and the ability to customize the security dashboards.

- **Security Application** - Metron provides standard SIEM like capabilities (alerting, threat intelligence framework, agents to ingest data sources) and packet replay utilities, evidence store and hunting services commonly used by SOC analysts.

- **Threat Intelligence Platform** - Metron provides next generation defense techniques, using a class of anomaly detection and machine learning algorithms that can be applied in real-time as events are streaming in.

At the beginning of the practical implementation of this thesis, we made an enormous endeavor to adopt Apache Hortonworks platform and specifically to deploy apache metron, this well known incubating Apache project. Hortonworks update its tutorials and guidelines for apache metron adoption and there is a live community around that project. However compelling Metron it may seem, even though we were truly fascinated by its encompassing nature with too many big data tools and applications, the deployment was quite devastating. After weeks of configurations, error checking and troubleshooting, Apache metron was decided to be abandoned. The reason behind our decision was that the demanding and intensive processing power needed to run apache metron surpassed the limits of our physical resources. Our virtual machine, was weak and couldn't afford the pressure needed for an even stripped down version of metron. After that, we considered metron too resource exhausting to utilize and too complicate to configure. Therefore, we continued our thesis with other tools.

9.4 Compelling Cyber Security Solutions or Complements

Sight Security, Inc., known as iSight Partners, Inc., provides cyber threat intelligence solutions to public sector and commercial entities in the United States and internationally. It offers ThreatScape suite, including ThreatScape Cyber Crime (financially motivated actors), Cyber Espionage (targeting corporate and government entities for strategic reasons), Hacktivism (politically or ideologically motivated threats), Critical Infrastructure (corporate and national infrastructure), Enterprise (geopolitical and legal trends affecting the business enterprise) and Vulnerability –Exploitation (vulnerabilities from discovery to patching, exploit development, regional propagation of exploit code and ongoing malicious campaigns in the wild). The company also provides ThreatService (a cyber intelligence service) including Threat Diagnostics (to identify an organization's threat profile highlighting threat sources targeting their assets

and associated tactical and strategic implications), Breach Diagnostics (to identify the scope of an incident and respond quickly to incidents) and intelligence integration services. In addition, it offers ThreatScape API enabling organizations to integrate the cyber threat intelligence with their security technologies and processes. The company was founded in 2006 and is based in Dallas, Texas. It has threat analysis centers in the United States, the Netherlands, India, and Australia. As of January 14, 2016, iSight Security, Inc. operates as a subsidiary of FireEye, Inc. **FireEye** is a cyber security company which provides software solutions for advanced and new cyber attacks, protecting its clients before, during and after a breach. In this context its products can identify connections between alerts, prioritize alerts and get actionable contextual intelligence for rapid remediation (Bloomberg, 2017).

On top of these collaborative and open sourced solutions about enhanced security, there is **AlienVault OTX** (OTX.Alienvault, 2017) a forum which provides open access to a global community of threat researchers and security professionals. The platform harnesses crowd-sourced wisdom, delivers community-generated threat data, enables collaborative research, and automates the process of updating your security infrastructure with threat data from multiple sources. Members can actively discuss, research, validate, and share the latest threat data, trends, and techniques. On one hand, AlienVault provides its widely used Open Source SIEM platform – (OSSIM) with ongoing development and on the other hand distributes its commercial Unified Security Management platform – (USM). OSSIM combines asset discovery, vulnerability assessment, intrusion detection, behavioral monitoring and SIEM in a single platform and it is completed with event collection, normalization and correlation in order to accelerate threat detection and compliance. Moreover, OTX data is integrated into OSSIM which can provide additional context to log data and security events as well as alarms, giving more visibility into weird activity. AlienVault Labs security research team publishes their threat research via their feed, but unfortunately only users of the commercial USM benefit from this out-of-the-box integration. Finally, users have the ability to export the Indicators of Compromise (IoCs) into several formats including OpenIOC, STIX, and csv, allowing them to be instrumented into user's security architecture (SIEM, access control devices). Alien Vault's IOCs include: IP addresses, Domains, Hostnames (subdomains), Email, URL, URI, File Hashes: MD5, SHA1, SHA256, PEHASH, IMPHASH, CIDR Rules, File Paths, MUTEX name and CVE number. Taking everything into account, Alien Vault offers shared security intelligence and increased infrastructure security visibility and network control.

Figure 15. Alien Vault's OTX logo

Martin Ussath et al, (Ussath,et al, 2016) consider Open Threat Exchange (OTX) from AlienVault a service that provides shared threat intelligence information. According to this research paper, users of OTX can create so-called pulses to share information with other community members. To create a pulse it is possible to upload PDF reports, plain text files or STIX documents. The parser of OTX tries to extract relevant indicators like IP addresses, domains, file hashes, file paths and email addresses. The shared information can be downloaded in different formats, such as CSV, OpenIOC 1.0 or 1.1 and STIX. Unfortunately, their brief evaluation of the OTX service showed that different indicators were extracted merged, because of the fact that the parser identified a product version number as an IP address, while the shared information taken from investigation reports or articles and threat intelligence information was not enriched with relevant context information.

Another Open Source solution is fetched by **IKANOW** enterprise. Its open source security analytics tool integrates with third-party applications and provides ingest, search, data widgets, and export features. The free Community Edition is a stripped-down version of the Enterprise Edition. It collects, stores, processes, retrieves, analyzes, and visualizes unstructured documents and structured records. Data from all sources is transformed into a single data model that allows common queries, scoring algorithms, and analytics to be applied across the entire dataset. Its open and flexible nature in conjunction with fast infoSec analytics enables Cyber-Security teams to contextualize breaches, prioritize vulnerabilities and characterize their threat posture while reducing enterprise risk. The open-source big data analytics platform of IKANOW is committed to liberating CISOs from the constraints of vendor lock-in and legacy technology architectures that impede improvements in information security posture. IKANOW enterprise has partnered with TI providers to gain valuable intelligence of customers by deepening insights into their overall security posture. It has integrated with iSight's ThreatScape feed to allow critical IOCs to quickly be ingested, extracted, and operationalized through automated historical lookups against network logs. The automated lookup process takes extracted IOCs from private threat feeds and open source blacklists and

generates alerts against historical log data. This fusion of threat feed and network log data saves time, money, and ensures that organizations have insight and knowledge of high threat activity on their networks. The IKANOW threat analytics platform also allows vulnerability and exploit information to be easily extracted from the ThreatScape feed, enabling an easy analyst pivot into enterprise scan data. IKANOW is applying adaptable analytical techniques to help mechanize the analysis and decision making process, resulting in lower cyber risks. Also it integrates with private TI feeds, social media, OSINT data, network logs, and enterprise data.

9.5 Batch vs Real – Time. A combat or a friendship?

As White (White, 2012) remarks: "MapReduce is fundamentally a batch processing system, and is not suitable for interactive analysis. You can't run a query and get results back in a few seconds or less. Queries typically take minutes or more, so it's best for offline use, where there isn't a human sitting in the processing loop waiting for results". However, the larger ecosystem of Hadoop, with many projects hosted by the Apache Software Foundation, has evolved beyond batch processing. Firstly appeared HBase, providing a vital solution for batch operations for reading and writing data in bulk, but as White continues : "the real enabler for new processing models in Hadoop was the introduction of YARN in Hadoop 2, a cluster resource management system, which allows any distributed program (not just MapReduce) to run on data in a Hadoop cluster". New processing patterns emerged like Impala or Hive on Tez, with which became possible to achieve low-latency responses for SQL queries on Hadoop. However, in iterative processing, such as those in machine learning, it's much more efficient to hold each intermediate working set in memory, in contrast to a disk. MapReduce does not allow this, but it's straightforward with other stream processing systems, enabling a highly exploratory style of working with datasets. White asserted that "Stream processing systems like Apache Storm, Spark Streaming, or Samza make it possible to run real-time, distributed computations on unbounded streams of data and emit results to Hadoop storage or external systems".

"The data streaming phenomenon was apparent at Strata + Hadoop World 2016 in San Jose, Calif., where Kafka and Spark Streaming often appeared in tandem in presentations that showed the latest activity on the Hadoop front lines" (Vaughan, 2016). According to TechTarget, Doug Cutting and Jay Kreps, key originators of Hadoop and Kafka respectively, were interviewed at the Strata conference, both discussing the ascent of streaming in big data applications. Doug Cutting, who is now chief architect at Hadoop distribution vendor Cloudera Inc., said that the original Hadoop implementation focused on data at rest. But with version 2.0 of Hadoop, released by The Apache Software Foundation in late 2013, big data streaming analytics became an important style of programming. The new forms of streaming software don't work at

the sub-millisecond rates achieved by some earlier products, however, are able to address large data sets at price-points far below predecessor systems. On the other hand, Kreps, who left a role as a principal staff engineer at LinkedIn in 2014 to co-found and be the CEO of Confluent Inc. (a startup offering a Kafka-based data streaming platform), remarked that users build streaming platforms collecting data from different parts of the organization for processing, and more specifically said "Streaming is a whole center of data for companies that adopt it". Finally, Vaughan claims that because both Kafka and Spark Streaming arise from open source projects, are capable of attracting an ever-growing cadre of skilled developers, whist other open source streaming frameworks, such as Apache Storm and, most recently, Apache Flink, auguring broader adoption of open source data streaming technology.

A major distinction between Batch and Real-time is latency. Below, the overall latency of a computation at a high level is comprised of communication network latency, computation latency, and database latency. The precise budget for each of the three broad categories (network, compute, and database) depends highly on the application. Compute-intensive applications will leave less room for network and database operations (CSA, 2014).

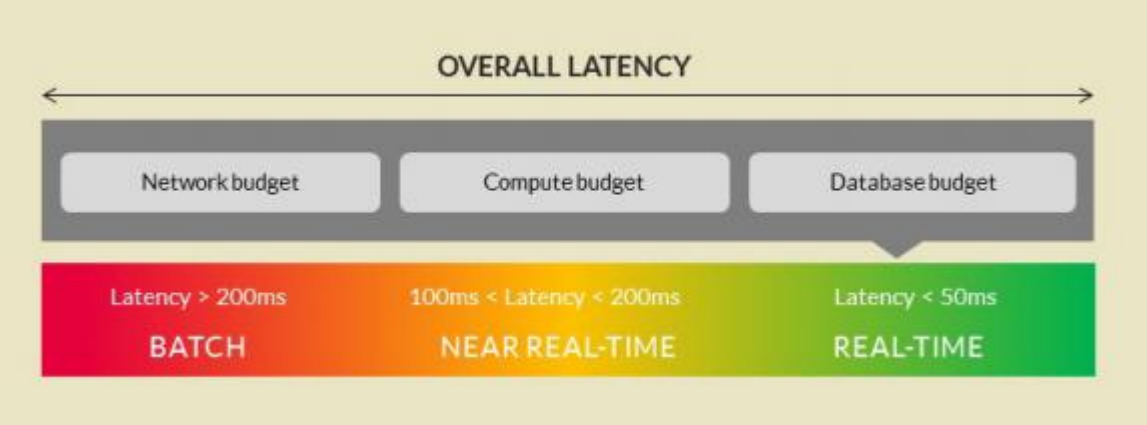
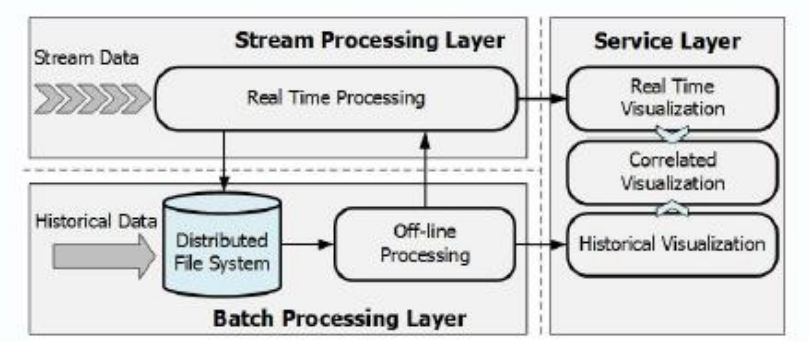


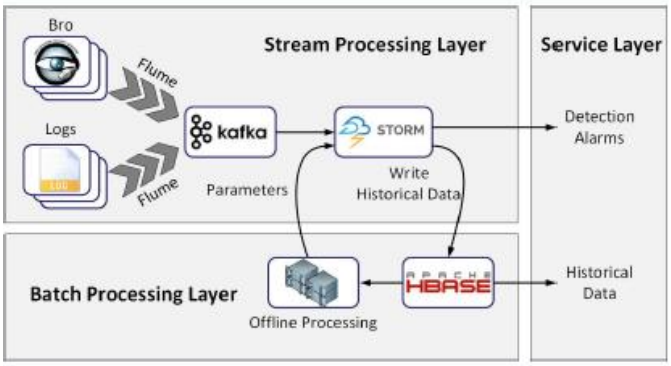
Figure 16. Overall Latency (CSA, 2014)

According to (Marz et al, 2015) both batch and stream paradigms, can be combined in the lambda architecture, to analyze big data in a real-time manner. Taking into account that batch processing produces high latency, with responses in the order of tens of seconds, while critical applications may require real-time processing, with responses within seconds; only stream processing techniques can reach such efficient levels and analyze swiftly massive unbounded data continuously generated.

A well balanced approach was presented by (Lobato et al, 2016) where their system was based on the lambda architecture, allowing real-time manipulation and analysis of massive amounts of information. The lambda architecture had three layers: the stream processing layer, the batch-processing layer, and the service layer.



The stream processing layer deals with the incoming data in real-time. The batch-processing layer analyses a huge amount of stored data in a distributed way through techniques such as map-reduce. Finally, the service layer combines the obtained information of the two previous layers to provide an output composed by analytic data to the user. Therefore, the lambda architecture goal is to analyze, accurately and in real-time, streaming data, even with its ever-changing incoming rate to obtain results in real-time based on historical data. The data analysis is divided in three steps: capture, normalization and processing. The system captures data, then, information is sent to the normalization process, data are formatted and enriched with external parameters, such as geographic information, correlations, and lastly, the normalized data are real –time processed, extracting security patterns. Once the system obtains the results, they are visualized and stored.



9.6 Streaming analytics frameworks for network monitoring

Gupta et al in their research (Gupta et al, 2016) combine network monitoring with stream processors, on the grounds that even though network operators need to collect voluminous and disparate data at extremely high rates (several terabits per second) with rich information about Security and network performance, yet efficiently analyzing them remains a challenge. They state that Network operators typically perform network management tasks while coping with fixed-function network monitoring capabilities, such as IPFIX and SNMP. Nevertheless, the advent of programmable hardware like OpenFlow

switches makes it possible not only to customize packet formats and protocols, but also to install custom monitoring capabilities in network devices that output data in formats that are amenable to scalable, distributed stream processing systems like Apache Spark or Apache Storm, which can perform streaming data analysis. Therefore, they argue that it may be possible to consider network monitoring as a stream processing problem, where each packet is represented by a tuple, and streams of packets comprise tuple streams for which many distributed stream processing programming idioms can apply. However, they claim that due to the inherently high rates of network traffic traversing a backbone network or a switch at a large Internet Exchange Point (IXP), realizing this programming abstraction requires reducing the traffic at the stream processor that does not satisfy the original query. So, they propose their prototype called Sonata which by partitioning of function between the switch and the stream processor; and with the ability to iteratively refine both the data plane rules for a query and its corresponding stream processing pipeline can reduce data rates at the stream processor by multiple orders of magnitude by pushing many of the filtering operations into the data plane. From their research we would like to maintain the idea that "*…the time is right to start thinking about how to apply streaming analytics frameworks to network monitoring. Doing so can ultimately help operators move from the current crippling set of technologies towards defining monitoring problems in terms of the questions they want to answer and the data they need to answer them*".

Evidently, streaming analytics frameworks can be deployed for network monitoring and many use cases of lambda architecture can be achieved utilizing whichever open source stream processing paradigm. Among **Apache Flink**, **Spark Streaming**, **Apache Samza** and **Apache Storm** we chose Storm, considering it reliable, scalable and easily deployable.

10 Apache Storm

According to tutorialspoint.com, Storm was originally created by Nathan Marz and his team at BackType social analytics company. Later, Storm was acquired and open-sourced by **Twitter**. In a short time, Apache Storm became a standard for distributed real-time processing system that allows processing of vast amount of data, similarly to Hadoop. Apache Storm is written in Java and Clojure programming languages and it is considered a leader in real-time analytics (tutorialspoint.com, 2017). Tutorialpoint.com is a website providing many basic tutorials, helping coders and ambitious internet users, to learn the basics of programming in many languages and study frameworks, libraries and undertake many more technical lessons.

Storm combined with the Twitter API, was the greatest example of real-time processing of millions of posts per second, yielding unprecedented insights at extremely high rates (Toshniwal, et al, 2014).

10.1 Main concepts of Storm:

Tuples. The main data structure in Storm. A tuple is a named list of values, where each value can be any type. Tuples are dynamically typed – the types of the fields do not need to be declared. Tuples have helper methods like getInteger and getString to get field values without having to cast the result. Tuples can contain integers, longs, shorts, bytes, strings, doubles, floats, booleans, and byte arrays. Also, custom serializers can be defined in order to include custom types into tuples. Usually we met n-tuples, with n represents the length of the tuple, thus the number of the values included, for instance (195.251.213.104) is a 4-tuple. Apparently, this 4-tuple could be an IP address or just a sequence of random integers. In tuples each value may responds to a String Attribute or an Integer Attribute, and thankfully getters of tuples render the access quite easy, so getInteger(), getIntegerByField(), getString() and getStringByField() are usual getters.

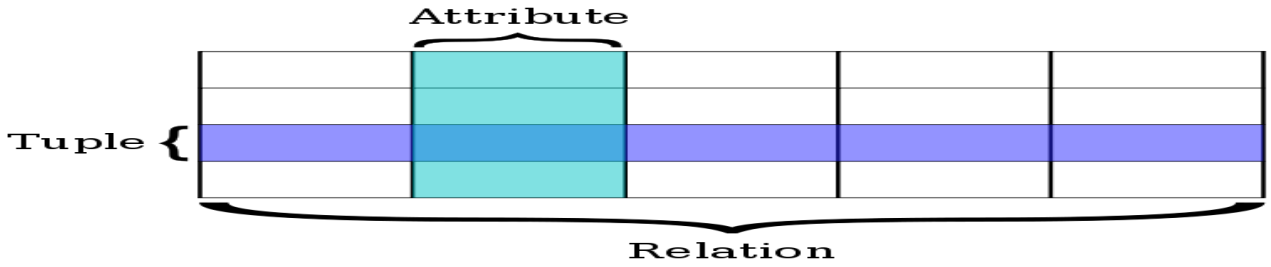


Figure 17. A tuple in Apache Storm

Necessary (security) stream data can be easily fit into tuples, just like (Gupta et al., 2016) did when demonstrated that each packet can be represented by a tuple with a collection of field tuple values corresponding to i) ts ii) locationID iii) sIP iv) sMac v) sPort vi) dIP vii) dMac viii) dPort ix) bytes x) payload. So, the tuple fields (packet) of their proposal included the timestamp – ts, the location of the packet in the network – locationID, the source IP, MacAddress, Port – sIP,sMac,sPort, the destination IP, MacAddress, Port – dIP,dMac,dPort, the total bytes of packet – bytes and the payload. Hence, obviously many similar security-related values could create a new tuple for processing.

Streams. An unbounded sequence of tuples, processed and created in a parallel distributed fashion. Every stream is given an id when declared, but since single-stream spouts and bolts are a commonplace the stream is may given the default id of "default".

Figure 18. A stream in Apache Storm

Spouts. A source of streams into topology. Usually spouts read tuples from an external source and emit them into the topology. Spouts are reliable (replay a tuple if it failed to be processed) or unreliable

(forget a tuple if it failed to be processed). Main method on spouts is nextTuple(), which either emits a new tuple into the topology or simply returns when there are no new tuples to emit. Attention is needed when implementing the nextTuple() method, due to the fact that Storm calls all the spout methods on the same thread. A Spout can emit more than one stream, as long as it is accurately configured through declareStream() and emit() methods (an interesting example is the spout reading from Twitter streaming API, attaining real-time processing of millions of posts per second).

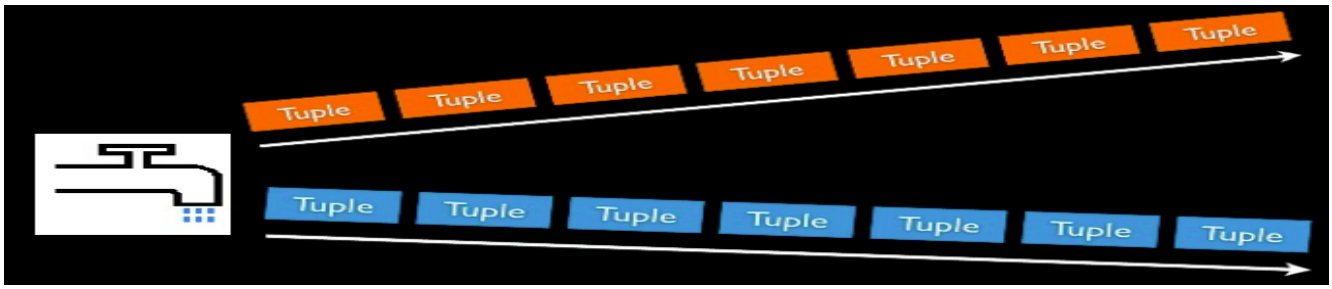


Figure 19. A spout in Apache Storm

Bolts. All processing in topologies is done in bolts. Bolts processes input streams and produces new streams Bolts can do anything from filtering, functions, aggregations, joins, talking to databases, or simple stream transformations. Complex transformations may require multiple bolts. A Bolt can emit more than one stream, as long as it is accurately configured through declareStream() and emit() methods. Bolts always subscribe to specific streams of another component and if you want to subscribe to all the streams of another component, you have to subscribe to each one individually. Main method in bolts is the execute method, taking as input a new tuple. Bolts emit new tuples, and call the ack method for every tuple they process so that Storm knows when tuples are completed. It's perfectly fine to launch new threads in bolts that do processing asynchronously.

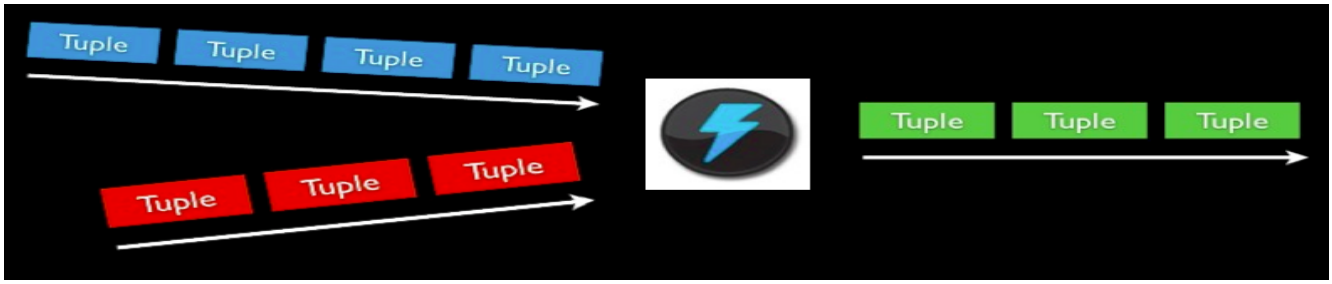


Figure 20. A bolt in Apache Storm

Topologies. A graph of Spouts and Bolts connected with stream groupings. Generally, a topology represents the whole picture of the real-time application. A topology runs forever, or at least until you kill

it, contrary to common practice in a MapReduce job. Topologies run either locally, mostly for testing, or on a production cluster.

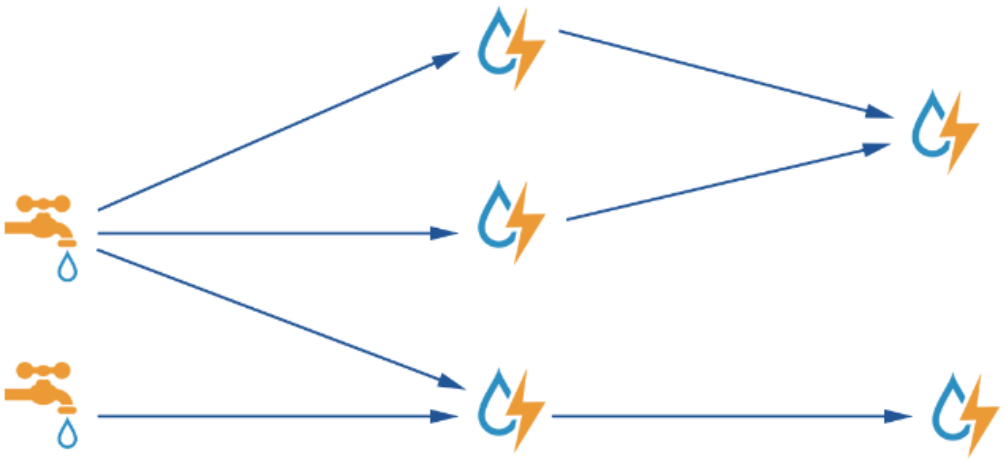


Figure 21. A topology in Apache Storm (Apache Storm, 2017)

***Stream groupings*.** Specify for each bolt which streams to receive as input. Stream groupings define how to send tuples from one set of tasks (thread) to another set of tasks (thread). Setting the parallelism hint for each spout or bolt in the setSpout() and setBolt() methods of TopologyBuilder. There are **8** built-in stream groupings in Storm, but we can implement custom stream groupings too. Specifically:

- **shuffleGrouping**: Tuples are randomly distributed across bolt's tasks and each bolt gets an equal number of tuples.
- **fieldsGrouping**: The stream is partitioned by the fields specified in the grouping (i.e. if the stream is grouped by the "userID" field, tuples with the same "userID" will always go to the same thread, but tuples with different "userID" may go to different threads.
- **Partial Key grouping**: The stream is partitioned by the fields specified in the grouping (like fieldsGrouping), but are load balanced between two downstream bolts, which provides better utilization of resources when the incoming data is skewed.
- **All grouping**: Stream is replicated across all bolt's threads. It should be used with care.
- **globalGrouping**: The entire stream goes to a single one of the bolt's thread, the thread with the lowest id.
- **None grouping**: This grouping specifies that you don't care how the stream is grouped. Currently, none groupings are equivalent to shuffle groupings, but eventually Storm will push down bolts with none groupings to execute in the same thread as the bolt or spout they subscribe from.

- **Direct grouping**: This is a special kind of grouping. A stream grouped this way means that the **producer** of the tuple decides which task of the consumer will receive this tuple. Direct groupings can only be declared on streams that have been declared as direct streams. Tuples emitted to a direct stream must be emitted using one of the emitDirect(int, int, java.util.List) methods. A bolt can get the task ids that the tuple was sent to.
- **Local or shuffle grouping**: If the target bolt has one or more threads in the same worker process, tuples will be shuffled to just those in-process tasks. Otherwise, it acts like a normal shuffle grouping.

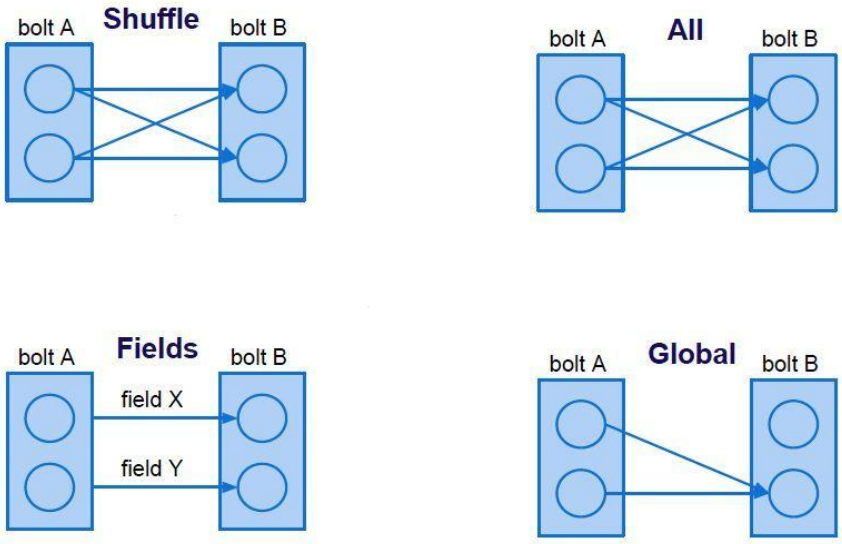


Figure 22. Common stream groupings (Leibiusky, J. 2012)

Reliability. Storm guarantees that every spout tuple will be fully processed by the topology. It does this by tracking the tree of tuples triggered by every spout tuple and determining when that tree of tuples has been successfully completed. Every topology has a "message timeout" associated with it. If Storm fails to detect that a spout tuple has been completed within that timeout, then it fails the tuple and replays it later. To take advantage of Storm's reliability capabilities, you must tell Storm when new edges in a tuple tree are being created and tell Storm whenever you've finished processing an individual tuple. These are done using the OutputCollector object that bolts use to emit tuples. Anchoring is done in the emit method, and you declare that you're finished with a tuple using the ack method.

10.2 What makes a Storm Topology

From the official website of Apache Storm we made some basic notes (Apache Storm, 2017). Firstly, we demystified the primary concept of a topology.

There are three main entities used to actually run a topology in a Storm cluster:

- **Worker processes**. A machine may run one or more worker processes. Each worker process is a physical JVM and belongs to a specific topology. In each worker process there may be numerous executors (threads) running. Worker processes are configured by the *setNumWorkers()* method (check configuration of a topology below – Figure 10).
- **Executors (threads).** Each executor is dedicated to a certain component, spout or bolt. In each executor there may be numerous tasks running. Executors (theads) are configured by the parallelism hint in setSpout() and setBolt() methods (check configuration of a topology below).
- **Tasks**. The task performs the actual data processing. Many tasks may be used for one component, spout or bolt. Tasks are configured by the setNumTasks() method (check configuration of a topology below). Beware that the setNumTasks() sets the sum of all tasks needed for that component, spanning either in one thread or more.

```
Config conf = new Config(); // basic configuration
conf.setNumWorkers(2); // use 2 worker processes
topologyBuilder.setSpout("a-spout", new aSpout(), 2); // set parallelism hint of "a-Spout" to 2
topologyBuilder.setBolt("b-bolt", new bBolt(), 2) // set parallelism hint of "b-Bolt" to 2
        .setNumTasks(4) // number of desired sum of tasks for that b-Bolt
        .shuffleGrouping("a-spout"); // tuples randomly & equally distributed to bolt
```

Finally, the **figure 24**, depicts a topology, wherein the sum of parallelism hints set the combined parallelism, a necessary value in order to estimate the number of executors (threads) spawned in each worker process. Basically, number of threads is yielded from:

Sum of all parallelism hints (executors) / number of desired worker processes

In other words:

Combined parallelism (executors) / number of worker processes

Respectively, the number of tasks inside a component (Spout or Bolt), is yielded from:

Tasks of a component / parallelism hint of the component (executors)

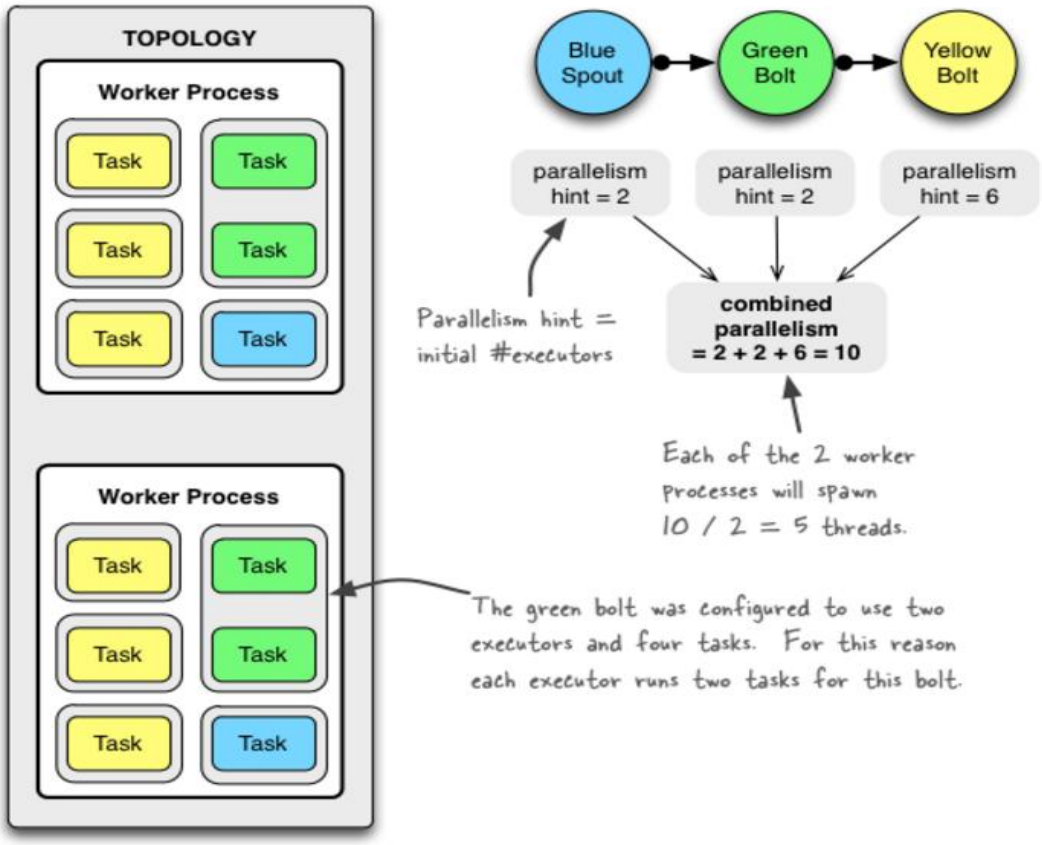


Figure 23. Basic Configuration of a Topology in Storm (Apache Storm, 2017)

Furthermore, there is a great feature in Storm, allowing increase or decrease of the number of worker processes and/or executors without being required to restart the cluster or the topology as illustrated below.

Reconfigure the topology "mytopology" to use 5 worker processes,

the spout "a-spout" to use 3 executors and the bolt "b-bolt" to use 10 executors.

$ storm rebalance mytopology -n 5 -e blue-spout=3 -e yellow-bolt=10

10.3 Spark vs Storm

Bobby Evans and Tom Graves from Yahoo, committers and Project Management Committee – (PMC) members to Apache Storm, Hadoop, Spark and Tez, made an early comparison between Spark and Storm, in order to outline appropriate use cases for each tool. To this end, they basically recommended these:

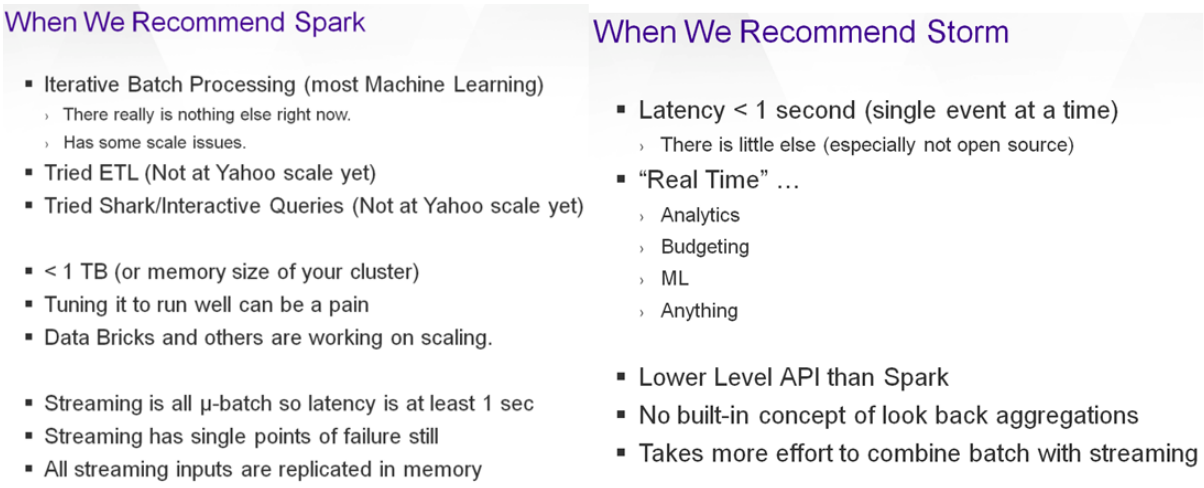


Figure 24. Spark 1.1 vs Storm 0.9.2 (Evans, B. & Graves, T., 2014).

Evidently, controversy sparked on whether Storm or Spark achieves greater results with stream data. However, P. Taylor Goetz, Apache Storm PMC Chair, member of technical staff at Hortonworks, claims Apache storm's superiority over Spark Streaming (Goetz, P. T. 2014). Even though admitting some bias towards Apache Storm, he considered inaccurate a number of comparing articles, due to the fact that configuration was almost always unavailable, therefore performance claims are unverifiable. Taylor Goetz points to comparative papers, wherein Spark Streaming (Micro-Batch) is compared to Core Storm (One-at-a-Time), although Storm's Trident (Micro-Batch) API would be a more appropriate comparator. He also admits, that Trident was a gradually escalating and ongoing developing feature in Storm, which in some comparative papers has not been taken into account, thus Spark seems to always outperform Storm. Taking a step back, Taylor remarks "Storm is a stream processing framework that also does micro-batching (Trident)", whereas "Spark is a batch processing framework that also does micro-batching (Spark Streaming)".

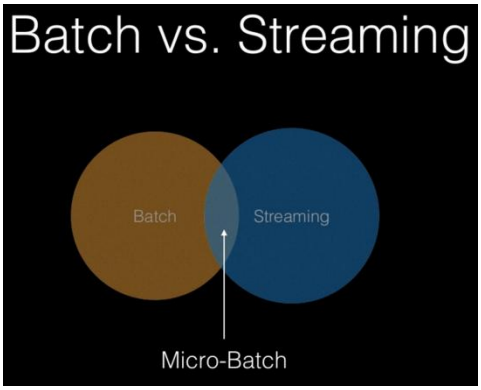


Figure 25. Micro-Batch (Goetz, P. T. 2014)

Taylor, after highlighting features of Apache Storm, like the 2 Streaming APIs: **A)** Core Storm (Spouts and Bolts, with low latency and operation on tuple streams and **B)** Trident (Streams and Operations), with

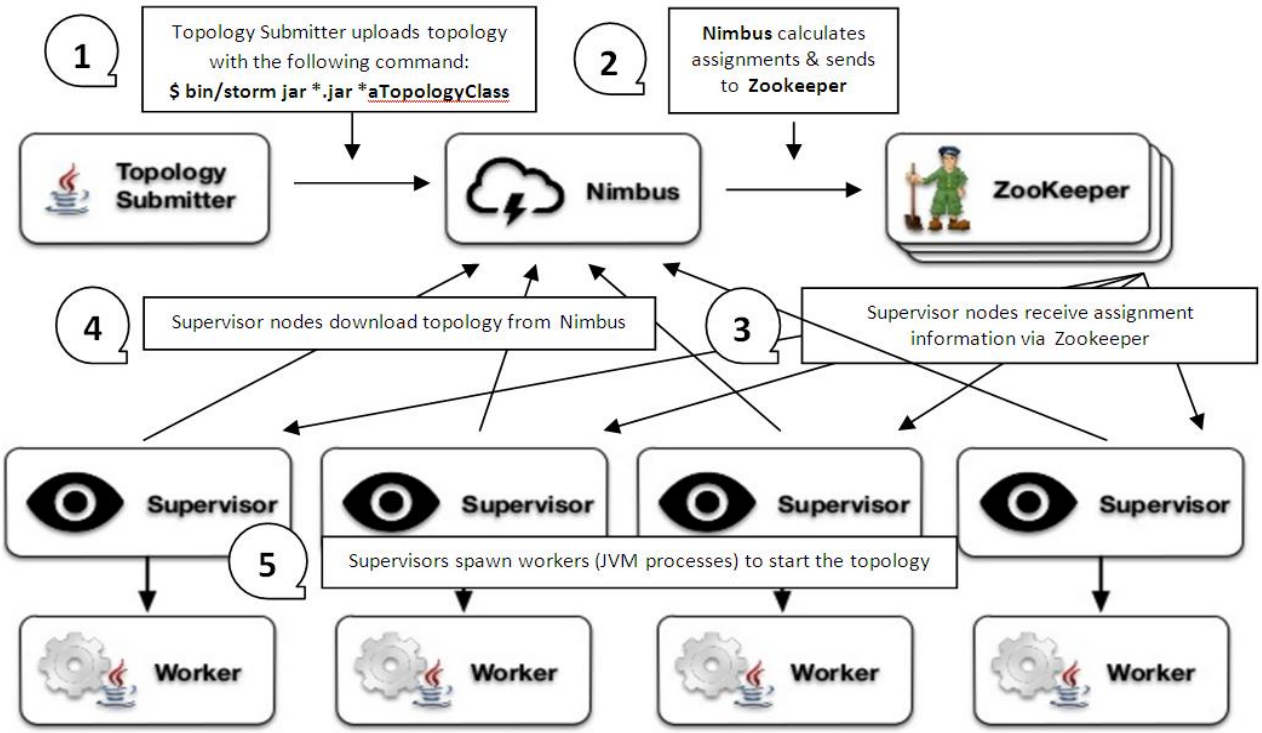
Micro-Batch, higher throughput and operation on streams of tuple batches and partitions, he examined the language options for both Storm and Spark, praising Storm's multi-language feature that allows the use of virtually any programming language.

Furthermore, P. Taylor Goetz scathingly emerged failure scenarios and mainly worker failures when comparing Storm with Spark. To this end, he grumble about Spark Streaming worker node failures, because if a worker node where a network receiver was running fails, then a tiny bit of data may be lost, that is, the data received by the system but not yet replicated to other node(s). Only HDFS-backed data sources are fully fault tolerant in Spark Streaming, however moving data into HDFS introduces additional latency. So, full fault tolerance still requires a data source that can replay data (e.g. Kafka)! On the other hand, worker failures in Apache Storm are confronted more precisely, because if a supervisor node fails, Nimbus will reassign that node's tasks to other nodes in the cluster and any tuples sent to a failed node will time out and be replayed (Similarly, in Trident, any batches will be replayed). Moreover he claims that with durable and reliable sources, Storm will not drop any data, while the ubiquitous pattern to attain an ideal source for Storm topologies involves Apache Kafka. However, Apache Storm output is well suited with many storage options, like Cassandra, HBase, HDFS, Kafka, Redis, Memcached, R, JMS, MongoDB and RDBMS.

Lastly, he evaluated performance of Trident API and Core AP with 3 supervisor nodes, which resulted in ~150k msg./sec. accompanied with ~80 ms latency for Core API and ~300k msg./sec. coming with ~250 ms latency for Trident API. Evidently, higher throughput comes with increased latency. Storm commercial use cases are numerous and are fully presented in the official Apache Storm website among which we can identify famous companies like Groupon, Yahoo, Spotify, Twitter, Alibaba, Baidu, Yelp and many more (Apache Storm, 2017).

10.4 Storm internals

In a holistic presentation on Real –Time Big Data with Apache Storm, P. Taylor Goetz, outlined basic concepts of Storm, diffusing rapidly the core architecture and the potential impact of real-time big data analytics with Apache Storm (Goetz, P. T. 2014). In conjunction with his in-depth presentation we created a timeline of submitting a Topology for clarity reasons, which we depict below:



10.4.1 Timeline of Submitting a Topology into Storm cluster

1. Topology Submitter uploads topology to Nimbus
2. Nimbus calculates assignments & sends to Zookeeper
3. Supervisor nodes receive assignment information via Zookeeper
4. Supervisor nodes download topology from Nimbus
5. Supervisors spawn workers (JVM processes) to start the topology

10.4.2 Heartbeats – Fault Tolerance

Furthermore, another great feature of Storm that we ought to highlight here is its fault tolerance. The way Storm establish fault tolerance is basically **heartbeats**. Specifically, workers heartbeat back to Supervisors and Nimbus via ZooKeeper, as well as locally to clarify their health status. So,

- If a worker dies (fails to heartbeat), the Supervisor will restart it
- If a worker dies repeatedly, Nimbus will reassign the work to other nodes in the cluster.
- If a supervisor node dies, Nimbus will reassign the work to other nodes.
- If Nimbus dies, topologies will continue to function normally, but won't be able to perform reassignments.

10.4.3 Reliable Processing

Finally, we outline the Reliable Processing feat of Storm. Reliable Processing includes the multiple anchorings which may form a tuple tree, backed by many bolts respectively. When a tuple has been processed successfully, the relevant Bolt can acknowledge this success, by delivering acks to Spout via a system-level bolt called "Acker Bolt". On the other hand, when a tuple has failed to be processed correctly, the corresponding Bolt can deliver a fail signal to Spout, in order to trigger the spout to replay the original tuple. Similarly an Acker Bolt is deployed. The same process is followed while processing any failure in the Tuple tree, which ultimately will trigger a replay of the original failed tuple.

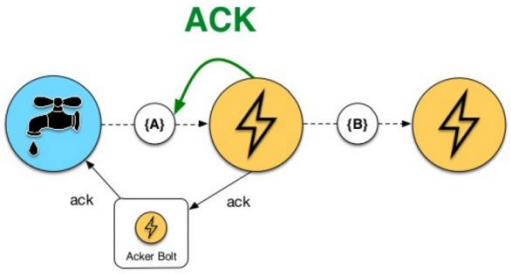


Figure 26. Delivering ack

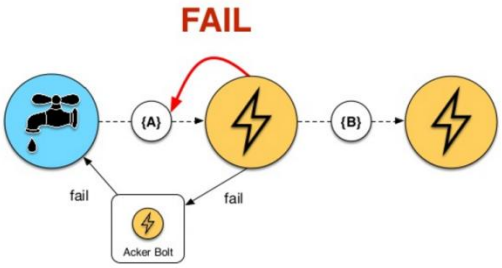


Figure 27. Delivering fail

11 Practical implementation of big data stream analytics

For our practical implementation, we beforehand set a clear goal; "***Accomplish a generally security enhanced solution utilizing a novel real-time stream processing tool combined with automatically updated cyber security intelligence***". Our application was deployed into Storm, therefore is real –time, fault tolerant, scalable and can claim all positive characteristics of Apache Storm.

To this end we were convinced that every endeavor to analyze stream data generated at high rates for security reasons would be helpful to prove our basic idea. Among numerous possibilities, we chose Apache Web Server, knowing that Apache was always a great solution regarding Web Servers and

maintained a praise-worthy reputation about its solid functionality, with a big market share among web server software providers.

Analyzing swiftly generated log files of an Apache Web Server, would be undoubtedly beneficial to our research. Analyzing basically meant to us, some form of real-time processing, to distinguish malicious from benign users. So, we decided to analyze each and every record yielded from an Apache web server, by parsing the whole record, extract the source IP and later, after creating our malicious – blacklisted dataset, to check if the incoming (source) IP was listed in our malicious dataset, in order to drop every packet originating from that user – potential attacker to our infrastructure.

Our application was influenced by the paper of (Lobato, et al, 2016). Lobato et al, adopted the lambda architecture and by utilizing kafka and apache storm they attained a real-time processing, normalization and visualization of results regarding malicious attackers. Our approach extends this system, by utilizing novel cyber security intelligence feeds throughout open APIs, bringing into out topology external mutually exchanged shared open security knowledge, in a timely manner to fortify our infrastructure.

Finally, we should declare that the whole experimentation and all topologies were run on a Vsphere Virtual Machine with 32GB RAM and 24 cores with installed the OS of Debian Jessie, a Web Server residing into the University of Macedonia. All testings were remotely executed, via a SSH connection to that web server, and the whole activities were basically deployed using Debian CLI. However, a graphical interface was installed (xfce) and sometimes used, mainly for the storm user interface, after establishing the secure SSH connection (with Public Key Infrastructure) and passing all the desired data through that confidential connection. The server was constantly monitored for attacks and other issues, and it worth noting that the amount of malicious users (source IPs) trying to establish a SSH connection to our system were interestingly too many (about a thousand SYN packets daily). Geolocating these IPs, China won the prize!

In order to examine and measure physical resources' capacity and utilization we used the htop tool, which is an easy to find tool on internet for all linux distributions. The interface below, is invoked with the command $ htop, and at the top we can chesck immediately CPU cores and RAM utilization, running processes and threads. Here we present the resources' utilization when our system was just booted, stripped from graphical interfaces. Only 540MB of RAM and 36 threads are used at an idle state.

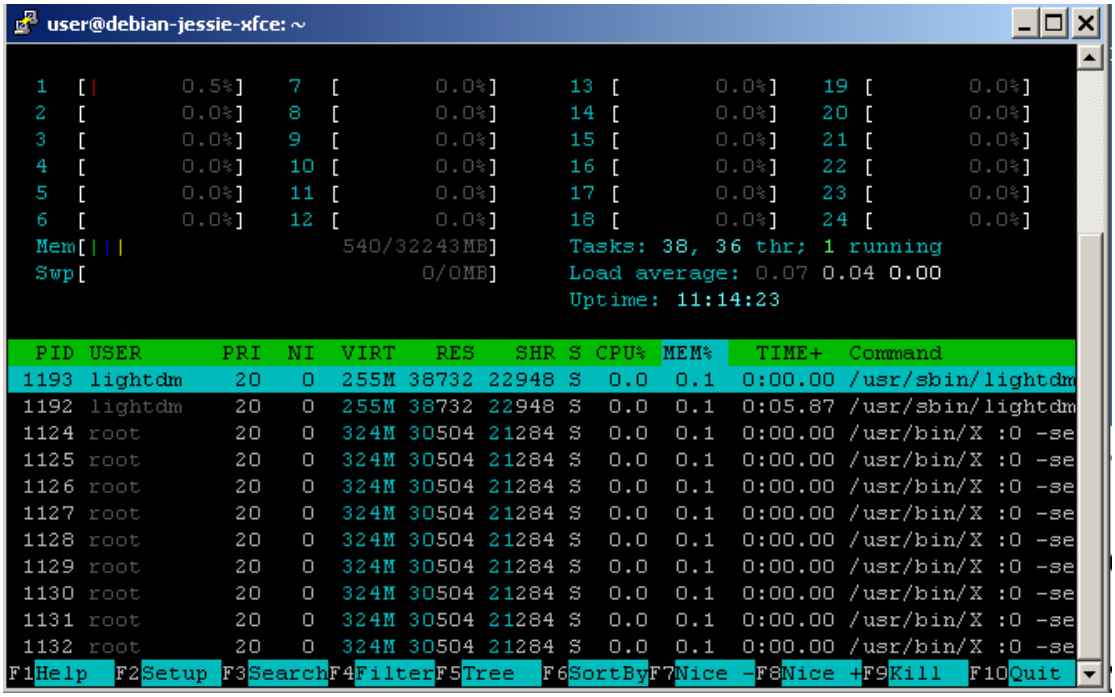


Figure 28. Our System's idle state (htop)

11.1 Step-by-step approach

Firstly we had to create our test data set (log file), similarly to an Apache Web Server, resembling exactly the generated log records. Thus, we created a tiny java application with a simple loop, to be able to compile our testing log file, filled with apache-web-server-like log records. We chose to create a 50 million (50.000.000) records and place it inside a proper place for log files (/var/logs/ourTesting…).

To visualize a similar record, Apache web server log files resembles this: '*85.202.20.1 - frank [10/Oct/2000:13:55:36 -0700] "GET /apache_pb.gif HTTP/1.0" 200 2326 "http://www.example.com/start.html" "Mozilla/4.08 [en] (Win98; I ;Nav)"*' (Apache website 2017)

Having our testing dataset ready, we had to create and deploy a storm topology to attain a big **data real time security analytics** paradigm. To this end, we hard coded our Spouts and Bolts to provide the desired functionality. Specifically, we populated our topology with three Spouts and two bolts; an IncomingSpout, a WebSpout, a ReportSpout, a ParseBolt and a Compare Bolt. Incoming Spout was just reading a stream of data from our previously created log file with millions of records, bringing it into our topology with the form of a tuple. Web Spout was the greatest manifestation of utilizing open source APIs to obtain almost real – time security intelligence feeds, and pass this intelligence into our topology, extending its security capabilities and fortify our system. Report Spout was an endeavor to report the results of a locally submitted topology in a timely manner, presenting the state of analysis to the user of the topology, using simple javascript functions into a Report.html document. However Report Spout was only used to submit

the topology locally and later was discarded, because Storm provides its own user interface. Parse bolt was the steam engine of our topology playing the vital role of parsing the log file to extract source IPs and pass it to the next bolt. Compare Bolt undertook the strenuous task to compare each and every tuple (source IP) with a java HashSet of malicious – blacklisted IPs and to flag the matching IP, compiling a second list of identified possibly attacking IPs and their corresponding trials throughout the extended life of our topology.

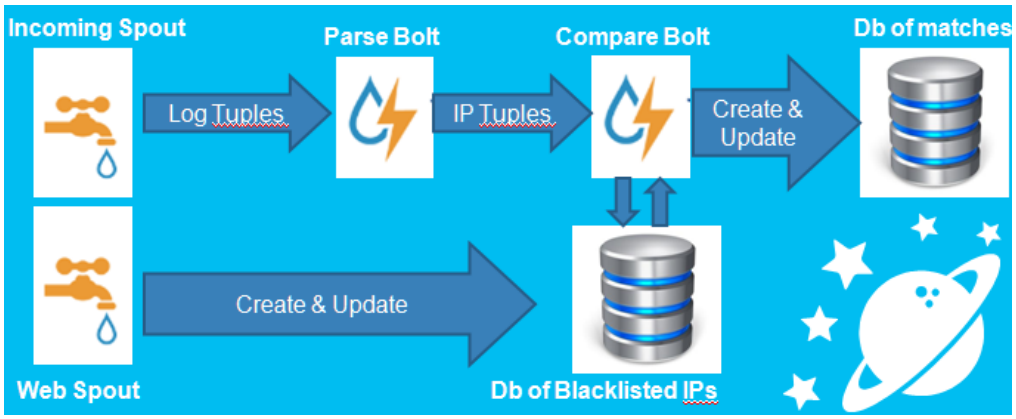


Figure 29. My Topology

To test the local submitter solution of Storm, firstly we deployed the topology "locally" without submitting it directly to nimbus. Unfortunately, without submitting it to nimbus our topology couldn't be visualized or checked for its state. That was the reason behind creating the third spout, called Report Spout, which was discarded later on, when our topology was submitted permanently to nimbus. This Spout was a small java implementation, with the task to periodically provide insights of the locally submitted topology. So, we our Report.html was auto updating to provide the soaring number of log records processed, the total number of currently blasklisted IPs, the updates of the database of blacklisted IPs, the number of identified potential attacks and the whole java HashMap of identified blacklisted IPs and its corresponding attempts to GET content from our web server.

After debugging our topology we run our testings numerous times locally and we observed a bottleneck when storm approached the number of 17 million records processed. Till then the rate of processing was great, approaching 17 millions of tuples into approximately 2 and a half minutes (159050 milliseconds), but after that point the efficiency was constantly deteriorating. So, we managed to deploy different topologies, retaining the basic concepts, increasing parallelism hints and/or number of tasks, however the results were quite the same.

Storm Topology - Our Results!

From totally **17327031 in 159050 seconds** processed Apache log files converted into Storm tuples,
we created secondary tuples fetching only the corresponding IP address targetting our Web Server.
Among that amount of tuples we swiftly distinguished totally **14** potentially malicious attacks!!!
In order to construct our unique and automatically updated Blacklist Database,
we chose **2 open APIs** of well-known blacklisted IPs feeds and **1 OTX Alien Vault** pulse!
Our blacklist database was configured to be regularly updated and during our testings
was updated totally **0** times (updates were set to occur every hour - 3600 seconds).
The number of blacklisted IPs populated into our Blacklist database was on average **9062**.
Finally our Map of malicious IPs to their corresponding attempts to establish connection with our system was registered into a HashMap.
Here we present the created HashMap too:
{109.121.167.229=1, 191.250.51.107=1, 131.161.9.253=1, 190.9.57.159=1, 85.157.119.115=1, 217.92.86.109=1, 201.52.218.97=1, 117.178.160.241=1, 74.120.91.148=1, 176.193.42.29=1, 185.118.154.54=1, 203.91.112.43=1, 63.243.252.179=1, 69.210.226.66=1}

Figure 30. Our Report.html (with Javascript functions)

Furthermore, the approach of locally submitting our topology had some side effects that should be referenced here. Primarily our local deployment was extremely memory intensive, utilizing up to 9196MB out of 32243MB of RAM. Also about 16 out of 24 cores were almost fully occupied and about 303 threads were spawned into our system to cope with the real-time processing.

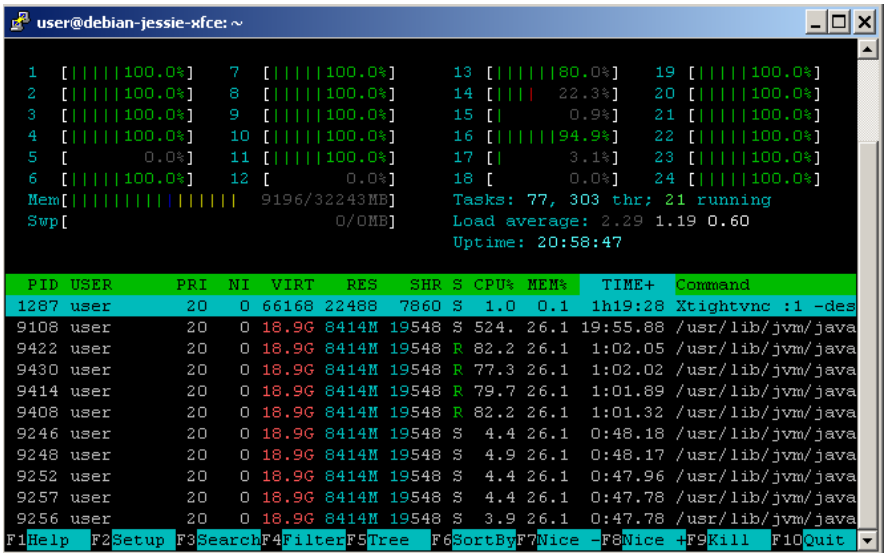


Figure 31. Resource intensive local topology (htop)

Afterwards, we proceeded to a regular topology submitter. We followed the strict rules of Apache Storm 1.1.0 documentation, we installed Zookeeper and configured our system (Zookeeper servers, nimbus host, ports and number of workers) and we awaked the Storm Nimbus, Storm Supervisor and Storm User Interface with the necessary commands.

$ zkServer.sh start

$ storm nimbus

$ storm supervisor

$ storm ui

After, submitting our Topology (our testing topology was initially called gTop Topology), we observed our results through storm interface.

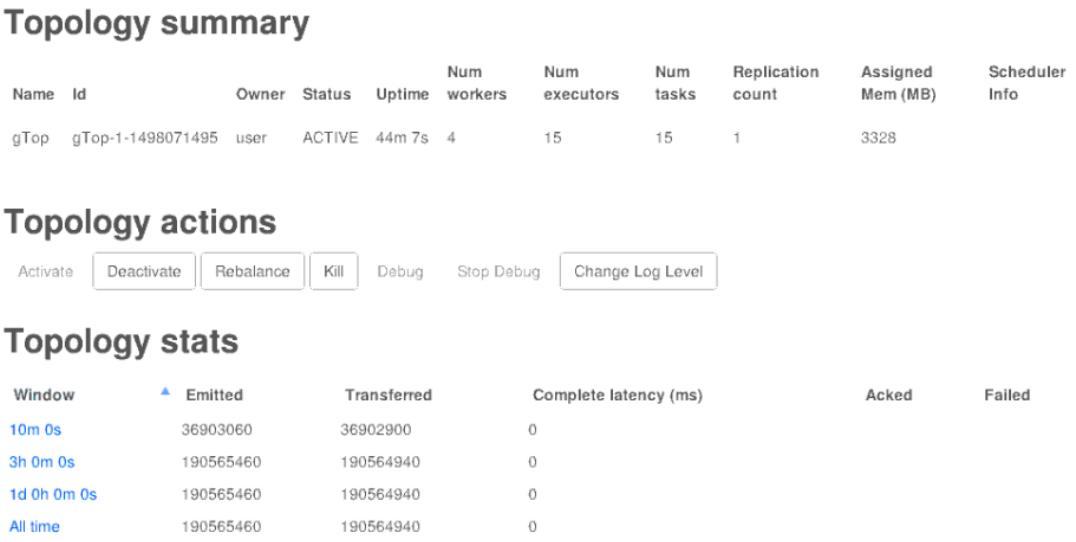


Figure 32. gTop topology (Storm UI)

As pictures present, our topology left running for some minutes (44 minutes) to examine its total rate of processing, knowing that at the beginning of the submitting, some configurations need some time to take place and start the real – time processing.

We can observe that roughly 36.903.060 tuples were processed at the last ten minutes, whereas our topology utilizing 4 worker processes and totally 15 threads with corresponding tasks. To explain that figures, we configured our topology to be able to spawn 4 worker processes and we defined the parallelism hint of spouts to 1, while the same hint was set to 4 for bolts. Therefore all together were 11. If we add the 4 ackers that were utilised to make our processing reliable and replay any missed tuples we reach 15 threads. The tasks were left blank, so similar amount of tasks were spawned.

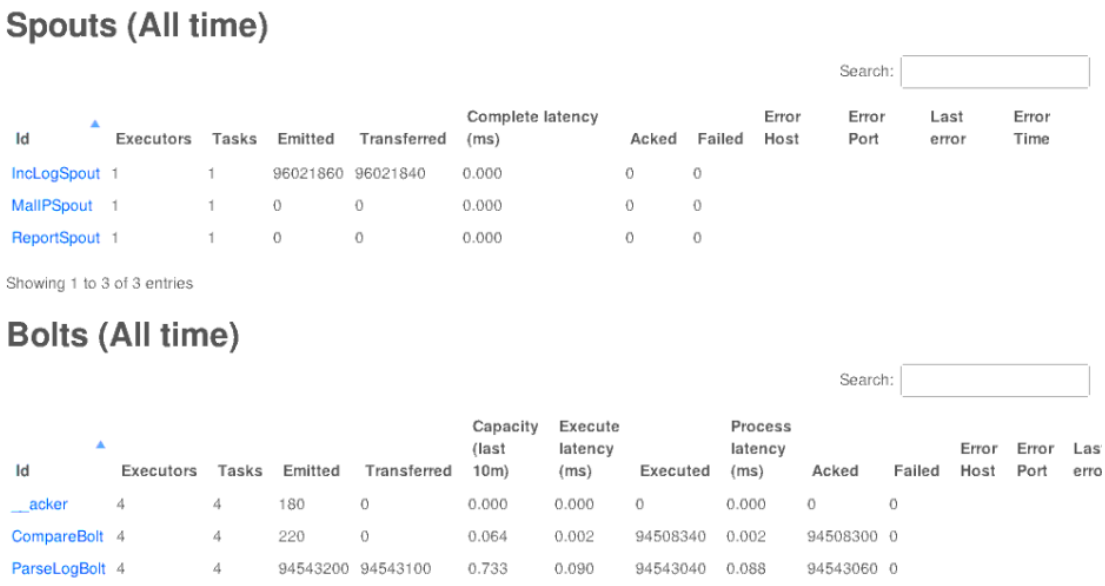


Figure 33. gTop's Spouts & Bolts (Storm UI)

It is clear that CompareBolt emitted roughly 220 tuples, which comprise the current number of malicious IPs detected to hit our resources and therefore could be later dropped or registered to be analyzed further or just be under surveillance by administrators.

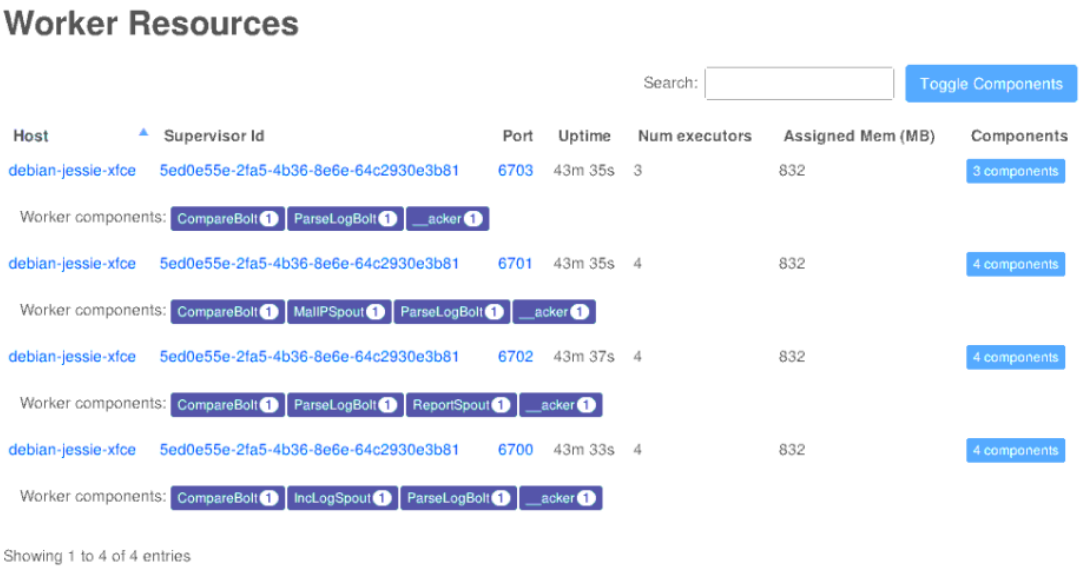


Figure 34. gTop's Worker Resources (Storm UI)

The names of Spouts and Bolts are clear and the number of their emmited tuples. We should focus on Capacity of ParseLogBolt, which sometimes exceeded the number of 1. That was a clear sign that the more demanding task was the task of ParseLogBolt, which strived to overcome the high amount of data. The capacity indicator is estimated by multiplying number of executed tuples with execute latency and dividing all by measurement time. Considering that ParseLogBolt was under strenuous conditions we had

to find a better deployment. The number of emitted tuples from CompareBolt are the desired tuples, carrying the malicious IPs that tried to establish connection with our system.

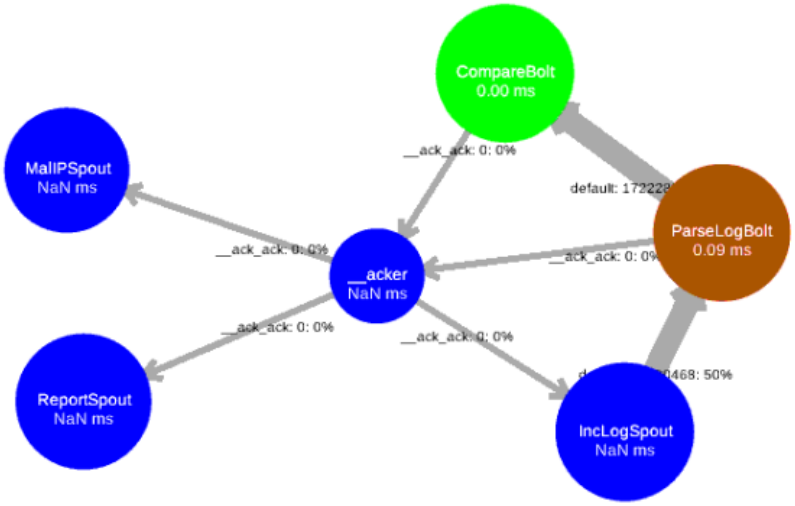


Figure 35. gTop's Visualization (Storm UI)

The directed acyclic graph (DAG), depicts our topology with its Spouts and Bolts, as well as the default ackers for reliable processing. We should note that the only specifically added ack took place inside CompareBolt, where we configured it to replay every missed corresponding initial Spout tuple.

However, that topology (gTop) was less memory and CPU core intensive (6646MB and almost 14 cores) but was quite impressive the number of spawned threads to deal with this situtation.

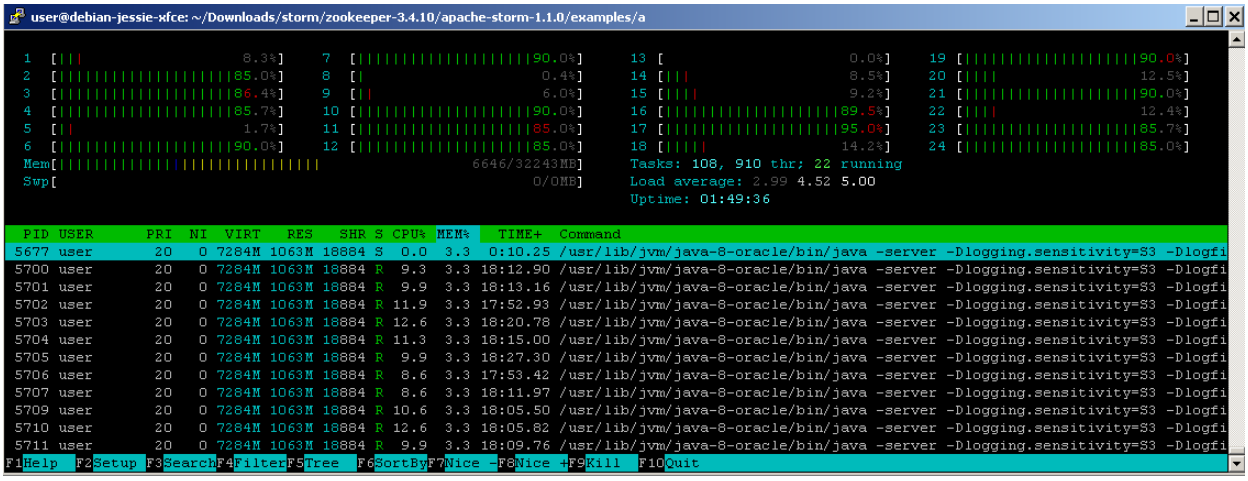


Figure 36. gTop's Resource Utilization (htop)

Taking into account the stressed bolt, we tweaked our Topology, creating a new one called hTop in which we gave ParseBolt some space to breath, by substantially increasing the number of relevant threads to 12. The results were refined and the bolt was now in normal activity. Moreover we stripped down our topology to merely dispensable parts; we emitted ReportSpout implementation, on the grounds that Storm provides its User Interface for submitted topologies.

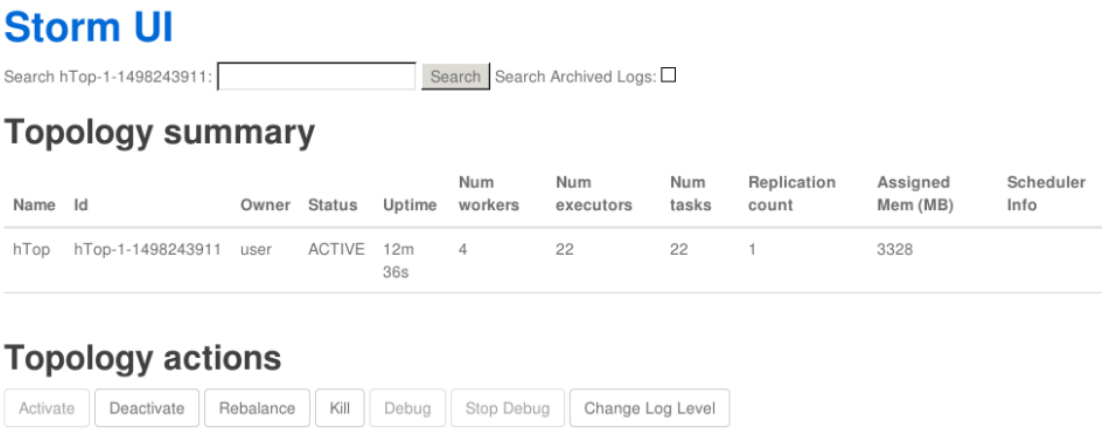


Figure 37. hTop Topology (Strom UI)

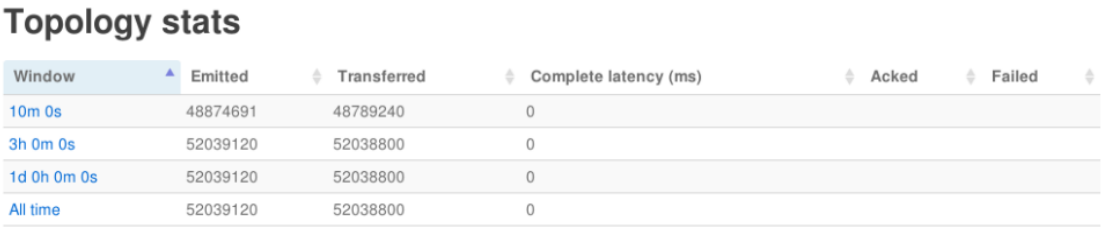


Figure 38. hTop's Stats (Storm UI)

Spouts (All time)

Search:

Id	Executors	Tasks	Emitted	Transferred	Complete latency (ms)	Acked	Failed	Error Host	Error Port	Last error	Error Time
IncLogSpout	1	1	26020240	26020240	0.000	0	0				
MalIPSpout	1	1	0	0	0.000	0	0				

Showing 1 to 2 of 2 entries

Bolts (All time)

Search:

Id	Executors	Tasks	Emitted	Transferred	Capacity (last 10m)	Execute latency (ms)	Executed	Process latency (ms)	Acked	Failed	Error Host	Error Port	Las erro
__acker	4	4	80	0	0.000	0.000	0	0.000	0	0			
CompareBolt	4	4	80	0	0.010	0.001	26017760	0.002	26017740	0			
ParseLogBolt	12	12	26018720	26018560	0.437	0.097	26018580	0.084	26018640	0			

Showing 1 to 3 of 3 entries

Figure 39. hTop Spouts & Bolts (Storm UI)

Here CompareBolt emitted roughly 80 tuples, which now comprise the current number of malicious IPs detected to have hit our resources and therefore could be later dropped or registered to be analyzed further or just be under surveillance by network administrators.

Worker Resources

Search: Toggle Components

Host	Supervisor Id	Port	Uptime	Num executors	Assigned Mem (MB)	Components
debian-jessie-xfce	5ed0e55e-2fa5-4b36-8e6e-64c2930e3b81	6707	5m 18s	5	832	3 components
Worker components:	CompareBolt 1 · ParseLogBolt 3 · __acker 1					
debian-jessie-xfce	5ed0e55e-2fa5-4b36-8e6e-64c2930e3b81	6705	5m 16s	6	832	4 components
Worker components:	CompareBolt 1 · MalIPSpout 1 · ParseLogBolt 3 · __acker 1					
debian-jessie-xfce	5ed0e55e-2fa5-4b36-8e6e-64c2930e3b81	6706	5m 15s	5	832	3 components
Worker components:	CompareBolt 1 · ParseLogBolt 3 · __acker 1					
debian-jessie-xfce	5ed0e55e-2fa5-4b36-8e6e-64c2930e3b81	6704	5m 16s	6	832	4 components
Worker components:	CompareBolt 1 · IncLogSpout 1 · ParseLogBolt 3 · __acker 1					

Showing 1 to 4 of 4 entries

Figure 40. hTop's Worker Resources (Strom UI)

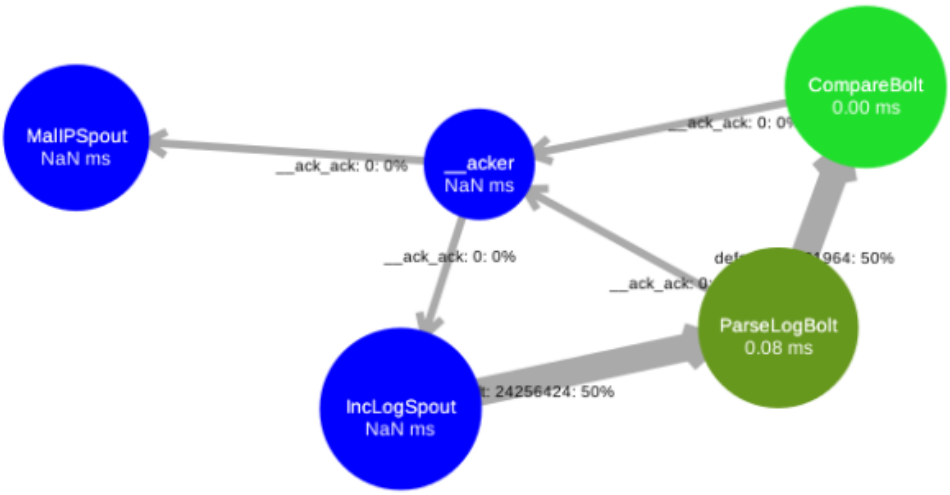
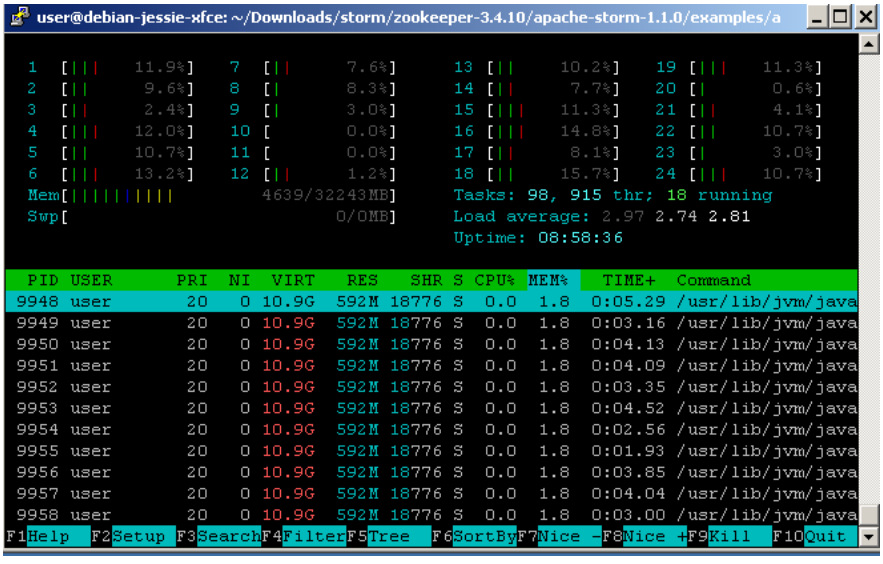


Figure 41. hTop's DAG (Storm UI)

After these simple tweaks, the total number of processed tuples were increased to amlost 41,5 million tuples per 10 minutes and the capacity of ParseLogBolt was remediated, decreasing to almost half value. To our hTop topology we retain the parallelism hint of 4, but we increased the threads of ParseLogBolt to 4, thus 12 executors – threads were spawned to underpin our previously stressed bolt. The number of emitted malicious IPs were roughly 20 from about 27,5 million processed tuples.



Finally we should mention that except applying more threads (via parallelism hint) or more tasks (via setNumTasks method) we can configure apache storm to delegate up to whatever number of worker processes we like.Therefore, we lastly tried to configure 8 worker processes into our system to check how strenuous may be. Well that configuration almost fully occupied the resources of our testing machine and

the system became unresponsive. CPU cores were exhausted and dangerously heated while 1273 threads were spawned to assist our last topology. We do not consider that configuration a viable example for our system so we didn't made more tests on that. Below is the utilization of our resources during the last attempt, in which we doubled our worker processes.

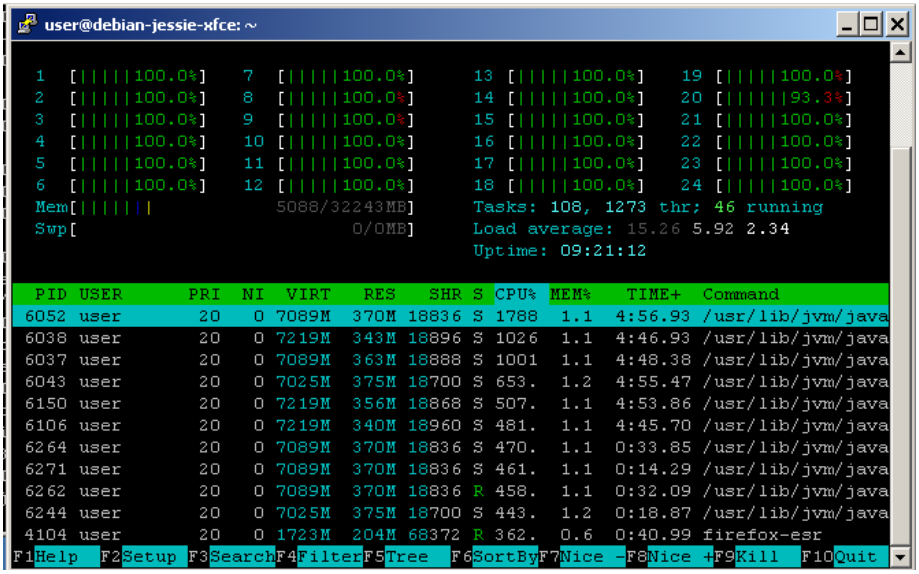


Figure 42. Double Worker Processes (htop)

11.2 Example Applications of my Storm deployment

In this section, we show how three network monitoring problems—reflection attack monitoring, application performance analysis, and port scan detection—can be expressed as streaming analytics problems.

Our approach regarding real –time stream processing of big data for enhanced Security may have numerous applications. Therefore here we can briefly enumerate some basic examples; however this list is not at all exhaustive.

11.2.1 Blocking Blacklisted IPs – Analysis of an attacker

This application resembles basically firewalls. Similar to packet filters' function, we can deploy a Storm Topology for real –time processing of high rates of incoming requests to a web server, in order to timely and without latencies distinguish malicious IPs from benign IPs requesting access to a website or web application and avoid establishing a TCP connection with potentially malicious users. The feed of malicious IPs can be extracted from various sources, like open APIs, public feeds or subscripted Security Intelligence feeds, which provide in regular intervals, ongoing blacklisted IPs. Moreover security intelligence feeds can even be created from internal processing of incoming packets which may present some potentially malicious activity.

Therefore, a deployed storm topology may discriminate between various packets or requests and swiftly drop any possibly malicious IPs trying to establish a TCP connection to our Web Server, while simultaneously retain a list of malicious IPs striving to gain access to our resources and process further the malicious requests to obtain security intelligence about the attacker or malicious user.

11.2.2 DNS traffic analysis – Confront DNS poisoning

Taking into account that "DNS is widely abused by Internet criminals" and that "DNS is used by cyber criminals in order to provide stealthy, flexible and resilient communication within malicious communication infrastructure" (Stevanovic et al, 2016), we should mention that DNS traffic abused for malicious purposes is commonly referred to as malicious DNS traffic. "DNS can facilitate stealthy and undisturbed communication as DNS traffic is present in all networks and it is not usually filtered by firewalls" (), therefore stream processing could be deployed for real –time processing of high rate DNS queries and responses in large scale operational networks.

Breaking down and maybe oversimplifying the DNS query process, we consider that the client's stub resolver after consulting local cashed records (like /etc/hosts, ldap, nis+, etc.) sends via the network a recursive DNS query to a (full DNS resolver) recursive DNS server (RDNS) which again after consulting its cache (if it is also a DNS cache server) and then it performs the desired recursion by which it discovers iteratively Authoritative Name Servers (ANS) for each zone and ultimately produces a mapping between the queried domain and the resolving IP address.

During this process we can deploy a Storm Topology for real – time processing of high rates of DNS queries and replies, in order to timely and without latencies distinguish malicious IPs and/or domains from benign IPs and/or domains and avoid originating a connection to a potentially malicious resource or more notably to prevent an unknown and obtrusive locally running malware to establish connection to its Command and Control - C&C Servers and extend its botnet, rendering our assets fully vulnerable to further exploitation. Again the feed of malicious IPs and/or domains can be extracted from various sources, like open APIs, public feeds or subscripted Security Intelligence feeds, which provide in regular intervals ongoing notorious - C&C Servers' IPs and/or domains. Again security intelligence feeds can even be created from internal processing of DNS queries and corresponding responses which may present some potentially malicious pattern.

Therefore, a deployed storm topology may set aside requested malicious domains or replied malicious IPs swiftly inform the user and/or network administrator about the potential harms of accessing the particular IP and/or domain, while simultaneously creating a log file about the exact procedure to understand the occurring attack i.e. either a benign domain was requested and a DNS Poisoning occurred

with falsified DNS reply or a malicious domain was requested probably by an already entrenched malware on our resources. Further process of the weird activity can take place to obtain significant security intelligence about the malware or the attacker.

11.2.3 Distributed port scanning detection.

Another interesting and ambitious deployment would be (distributed) port scanning detection. Existing intrusion prevention system (IPS) devices often cannot process traffic at high rates, and they typically only operate at a single network location. Instead, our topology could be tweaked to count the number of distinct SYN packets that never have a corresponding ACK packet (SYN scanning type) or FIN packets without previous corresponding SYN, ACK packets (FIN scanning type) as long as we are referring to a stateless firewall. We can convert each packet into a tuple, and Storm Bolts could couple each SYN with a matching ACK, if any, or retain a list of source IPs that sent a lone FIN. Many assumptions are made here, on the grounds that port scanning techniques should be carefully studied to apply the respective rules for each method. However, a Topology could easily play the role of a distributed real-time and fault-tolerant Intrusion Prevention System – IPS (or IDS) confronting with high rates of input data.

When referring to distributed port scanning our mind travels to Nmap uses a variety of active probing techniques and offers its users the ability to randomize destination IPs and change the order of and timing between packets. This functionality can obscure the port scanning activity and thus fool intrusion detection systems - IDS.

Non exhaustive list of examples

We remain confident that the aforementioned examples comprise a non exhaustive compiled list, regarding potential applications of Apache Storm (or other real-time processing paradigm) for enhanced security. To our knowledge practical use cases of Apache Storm can be encompassed into real-time IDS and IPS to optimize anomaly based detection implementations.

11.3 Notes on chosen Blacklist IPs datasets

Seeking a free trusted and regularly updated database of blacklisted IPs we stumbled upon the website of Delta Consultants (myip.ms), located in Commonwealth of Dominica (Caribbean), which ambitious goal is to help webmasters protect their websites from dangerous/scam IP addresses and spam bots. To attain their goal they created a real time blacklist of IP addresses based on statistics from their websites, which encounter tens of thousands of daily visitors, analysed for block spam bots and other threats by a simple protection system, which guarantees that only unknown crawlers/spam bots masking themselves as normal users are flagged excluding known search crawlers like Google Bot, Yahoo Bot, Bing Bot.

The chosen blacklist is called latest blacklist IP and includes all added IPs to Myip.ms blacklist Database during the last 10 days. Usually this list contains about 2,500 malicious IPs and the small time window of ten days ensures that only current spam bots and not valid users are enlisted. Blacklist text files are being updated daily every hour.

Moreover, we used another quite often referenced blacklist IP database (ipspamlist.com), provided by NoVirusThanks company. This list encompass malicious IP addresses engaged in hacking attempts, spam comments, postfix\imap scans, telnet scans, SSH brute force attacks FTP brute force, port scanning, postfix\email hacks, wordpress hacking and so on, identified by the company's honeypots and spam traps. This service is useful for threat intelligence and to help in the detection of malicious IP addresses.

These blacklists are two of many possible lists permeating the internet. Numerous open lists/feeds as well as closed – commercial ones can be found. The user can deploy only one or many lists, as long as duplicates are avoided to minimize processing overhead beforehand.

Regarding OTX Alien Vault pulses, the reader an\d/or researcher is free to choose his own pulse, subscribe to it and gain access to vital security intelligence. There is a wide variety of security intelligence exchanged in alien vault and indeed should be exhaustively searched to subscribe to the desired and pulse.
It is evident that the dataset of blacklisted IPs was tailor-made, on the grounds that we freely chose the appropriate blacklist feeds and we freely subscribed to relevant Alien Vault's pulses to incorporate such information into our topology. However, our system is scalable and every potential pulse or list could be also added to provide deeper examination of incoming IPs.

11.4 REFUTATION!

Apparently our approach presents some limitations. We clarify that these limitations indeed have been taken into account, however the whole premise of this research was constructed upon the notion that real – time big data security analytics can and should be applied utilizing appropriate novel tools to overcome traditional barriers and exceed the limits of current deployments.

We admit that however interesting it may seem, our "firewall" application may arise various concerns about its effectiveness, due to the fact that contemporary advanced cyber attacks deploy methods that bypass mere firewalls and simple malicious IP-based Access Control Lists (ACLs). Not to mention the IP spoofing techniques that render source IP useless information. Moreover we fully apprehend that blacklisted IPs are time delayed and may not provide a resilient and full protection against attacks. However we cherish the notion of the layered onion defense model (defense-in-depth) rather than the notion of lollipop defense model (Mavridis, 2015). It is clear that many sub methods can be undertaken to add a tiny grain to whole security infrastructure into a defense-at-depth security model.

But firstly we should declare our limitations setting our approach to a well-known timeline, and afterwards to dive briefly into two significant aspects of DNS traffic and DNS security extensions that nowadays almost fully protect us from some relevant attacks.

11.5 Limitations

11.5.1 Blacklist IPs

Our approach basically rotates around blacklisted IPs and notorious domains. We admit that feeds of blacklisted IPs are a delayed manner of acquiring knowledge of malicious IPs, therefore the attack and its using IP should be already flagged as malicious by another user or software. Apparently, brand-new malicious IPs that targeting our infrastructure could not be perceived and confronted by our proposal. The whole concept of blacklisted IPs involve mutual engagement and permeating attitude on sharing Security Information with each other timely and accurately. However many initiatives have been undertaken to make that possible and maybe gradually can present a viable alternative to closed commercial propositions. As stated earlier to our paper, OTX pulses, iSight Threatscape and other APIs with providing feeds of Security Intelligence is already implemented and widely accepted as accurate.

11.5.1.1 Locate our approach into zero day attack lifecycle

Here we would like to make a reference to zero day attacks and the lifecycle of vulnerabilities in order to understand the evolution of a scathing attack and sincerely locate our approach in that timeline.
Firtly, we noticed that the life cycle of a vulnerability presents a major security gap; which Frei called “known unknown vulnerability” (Frei, S., 2014)

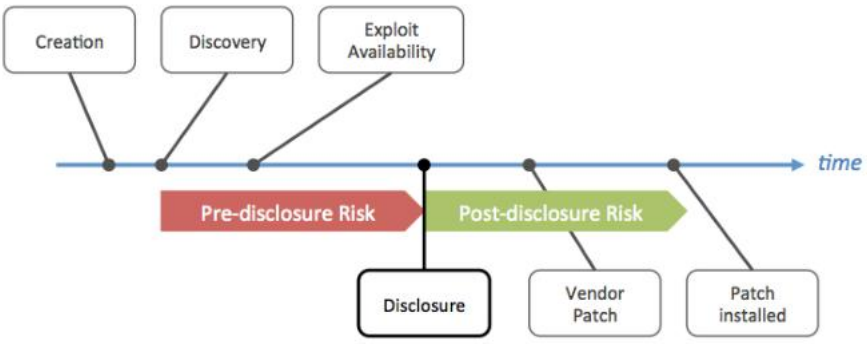


Figure 43.Life Cycle of a Vulnerability (Frei, S., 2014)

Frei explained the Security Ecosystem & Vulnerability Life Cycle, maintained that the life cycle of a vulnerability can be divided into phases between distinct events. According to his perspective, each phase reflects a specific state of the vulnerability and an associated risk exposure for the users of the affected

software. In order to capture these phases, six events were defined in the vulnerability life cycle: i) creation, ii) discovery, iii) exploit availability, iv) disclosure, v) patch availability, and vi) patch installation, as shown in his above Figure. Frei backs his words admitting that the exact sequence of these events varies among individual vulnerabilities and the parties involved. So, the two major phases are:

Pre-disclosure Phase; during this phase the public is not aware of the vulnerability and therefore cannot assess the potential risk or take any mitigating action. Within a privileged group, however, the vulnerability is known to exist; therefore, the vulnerability is regarded as a "**known unknown**" and can be potentially exploited by malicious participants. **Post-disclosure Phase;** this phase starts as soon as the public becomes aware of the disclosed vulnerability and the users are provided with information that will allow them to assess the risk or to take mitigating action until a patch is released and installed, thereby remediating the root cause of the vulnerability.

Frei frankly states that the exact sequence of these events is dependent on the manner in which the vulnerability information is managed by the discoverer, and as such, "it is a direct function of the incentives and ethics of the discoverer" (Frei, 2014).

After that, we consider that our approach is surely located after the exploit availability and in the best optimistic scenario almost matches the point of zero day attack. Apparently the first attempt to exploit a vulnerability coincide with the first deployed host to accommodate that attack and the so-called zero day of this attack. After the attack has caused harm or damage and this is undoubtedly identified, the attacking source IP, Port, Mac Address and many more network or other characteristics will be thoroughly analyzed by the victim and hopefully instantly disseminated in public available feeds for malicious IPs or IoCs and the like Security Intelligence feeds. However time-demanding it may seem, it is also a reality to our interconnected digital world to disseminate swiftly security intelligence all over the world within seconds, thus our approach remains an engaging way to partially protect our critical assets from quite novel attacks and we believe it adds a minor step to bridge security gaps. Finally, we should also highlight that Stevanovic et al, acknowledge that blacklisted IPs are indeed necessary to be utilized for modern security countermeasures, on the grounds that they encompass publicly blacklisted IPs to their proposal.

11.5.2 DNS queries

Furthermore, regarding the proposition of utilizing the same approach or processing method to DNS queries and replies again presents some limitations that are admitted here. Firstly, we should state that internet criminals can rely on either static hosting strategies or dynamic hosting strategies similar to content distribution networks (CDNs). "However, miscreants commonly adopt dynamic hosting strategies as they provide high availability and resilience against countermeasures. These strategies are characterized by

highly dynamic FQDNs-to-IPs mappings and are often referred to as agile DNS" (Stevanovic et al, 2016). As Stevanovic et al, argue, agile DNS strategies are typically categorized as Fast-flux (or IP-flux) and Domain-flux. Studying these strategies is easy to remember that IP-flux refers to a constant changing of IP addresses associated with a particular Fully Qualified Domain Name - FQDN, while Domain-flux refers to constant changing and allocation of multiple FQDNs to a single or multiple IP addresses. Either in fast-flux or in domain-flux, which already count some variations, attackers disguise their malicious domain and/or IP behind an extremely large pool of resources and an elusive short living. Therefore many conventional techniques to apprehend an attacker fade away. As Stevanovic et al, remarked recent studies indicate that Fast-flux strategies abused by modern botnets such as Storm are characterized with domain names hosted over more than 10.000 hosts, when modern and well known variations of the Domain-flux, commonly based on Domain Generation Algorithms (DGA), such as Conficker botnet which could generate 50.000 domains per day, from which 500 are queried daily (Antonakakis et al, 2012).

Taking that into consideration an approach that do not confront with these attacking strategies maybe is seen as outdated. However, Stevanovic et al, paved the way to escalate the adaptive DNSMap tool, which can label agile DNS traffic, and thus assumes a role similar to blacklisted IPs. DNSMap extracted mappings could be utilized to feed our engine. Nevertheless, established well-known approaches are already in place and are briefly but thoroughly presented by Stevanovic et al, at "Related work" section.

11.5.3 DNSSEC

Finally, we should mention the DNSSEC, which comprises a suite of specifications for securing certain kinds of information provided by DNS, providing origin authentication, authenticated denial of existence and data integrity. Utilizing DNSSEC with its digitally signed answers DNS cache poisoning is quite impossible. Today Google Public DNS is a freely provided public DNS service, fully supporting DNSSEC, while Verisign supports a free public DNS resolver service which performs DNSSEC validation. To our readers' interest Google DNS addresses are **8.8.8.8** and **8.8.4.4** while Verisign DNS addresses are **64.6.64.6** and **64.6.65.6**. (Google, 2017) & (Verisign, 2015). The exact implementation of DNSSEC is out of the scope of this research.

11.5.4 Port Scanning

Last but not least, we are aware of famous open source network analysis frameworks like Bro Project that are different from typical IDS and can play a pivotal role to the battle against cyber attacks and provide significant insights about network traffic, analyzing protocols and enabling high-level semantic analysis at the application layer. However, here among other features we praise bro about its port scanning detection. Specifically in Bro, network administrator (or just the user of Bro) can set the **time interval**

within failed connection attempts are tracked for the host or port detection (namely address scan interval or port scan interval) and also can configure the threshold of the number of unique hosts or ports a scanning host has to have failed connection attempts with a single port or host respectively (namely address scan threshold and port scan threshold) (Bro, 2017).

All in all, we have already outlined the benefits of real-time processing of voluminous data coming in extremely high rates and to this end many more variations of our approach could be implemented to chase a solution to many novel or traditional attacking methods. Nonetheless, Apache Storm presents unique desirable characteristics, like horizontal scalability and processing reliability.

12 Conclusions – Future Research

Taking everything into consideration, this thesis was a long journey among big data and security. After enumerated contemporary examples of swiftly generated vast amount of data each and every second, we reviewed the basic definitions of Big data, IoT and Cloud computing. Intertwined notions, yet distinguishing areas.

However compelling the novel paradigm may seem, security issues and challenges could immediately arise. So, a solid framework to study security was necessary; like RMIAS. These reference model have gained momentum among similar models and is widely accepted as secure and accurate enough.

Apparently, big data and IoT brought traditional security issues into light and extended further the perimeter needed security countermeasures. Interchangeably misused security notions were clarified, ongoing technological advancements were outlined and the way was paved to future promising deployments.

Significant challenges could keep some of us awake, when contemplating that even implantable sensors can be hacked. However ominous this technology may sound, can and should be used in favor of Security. On these grounds our Apache Storm deployment, was a basic manifestation of utilizing novel big data stream processing tools accompanied with cyber intelligence feeds to enhance our security countermeasures; gaining momentum of reliable, scalable and fault tolerant paradigms like Apache Storm. Possible applications of our approach were enumerated spanning from real-time blacklisting IPs, to confront DNS poisoning and/or port scanning techniques.

Our goal to outline contemporary cyber security notions combined with Big Data tools was achieved, and the reader can apprehend the enormous potential. Through our application we accomplished a generally security enhanced solution utilizing a real-time big data stream processing tool with cyber

security intelligence feeds. Public awareness should have been raised on contemporary security issues, and a novel processing example was outlined in detail to prove the enhanced security through big data tools.

12.1 Research Limitations

Our research was full of security challenges accompanied with promising results. We outlined both sides of novel technologies, even though limitations always arise and should be confronted. Our practical section in chapter 12, in which we deployed Apache Storm, was fully explanatory on limitations of each proposed example application. We declared our refutation and we believe that we have alleviated primary concerns on our limitations. We hope that limitations should not pose obstacles to researchers but motivate them to move further and exceed the entrenched limits.

12.2 Personal Challenges

During this thesis, many personal challenges emerged. If we would like to enumerate only the knowledge-based ones, we shouldn't exclude the clarification of complicated and overlapping contemporary security notions, distorted by marketers, the cryptographic algorithms and protocols negotiating a secure communication, the necessary programming languages and concepts (mainly JAVA) to realize every imaginary testing thought, the utilization of remote resources to attain your experiments safely and confidently, the OS elaboration to understand the differences and basically the similarities among them and finally the interaction with remote terminals and its dispensable award winning patience when networks or "unknown" issues occur.

12.3 Future Research

Undoubtedly, future research should be done on utilizing real-time stream analytics paradigms combined with machine learning techniques. Our ambition is to walk further down the academic avenue in order to obtain holistic knowledge on many relevant Security domains to provide much more in depth analysis and emerge the necessary big data tools to confront unseen zero day attacks. To this end, we consider machine learning algorithms and specifically implementing deep learning for enhanced Security our next step aligned to our personal goals.

References

Acquisto, D., G., Domingo-Ferrer, J., Kikiras, P., Torra, V., Montjoye, Y.A. & Bourka, A. (2015). Privacy by design in big data An overview of privacy enhancing technologies in the era of big data analytics. *European Union Agency for Network and Information Security - ENISA*. Retrieved from https://www.enisa.europa.eu/publications/big-data-protection/at_download/fullReport

Ahlm, E. & Litan, A. (2016). Market Trends: User and Entity Behavior Analytics Expand Their Market Reach. Retrieved from Gartner Research database (https://www.gartner.com/doc/3294335/market-trends-user-entity-behavior).

Ahlm, E. (2017). Roles and Responsibilities of Gartner's Analyst & Research Director, Eric Ahlm. *Gartner*. Retrieved from https://www.gartner.com/analyst/40943

Antonakakis, M., Perdisci, R., Nadji, Y., Vasiloglou II, N., Abu-Nimeh, S., Lee, W., & Dagon, D. (2012, August). From Throw-Away Traffic to Bots: Detecting the Rise of DGA-Based Malware. In *USENIX security symposium* (Vol. 12).

Apache Storm Official Website (2017). Apache Storm. *Apache Software Foundation*. Retrieved on May 2017 from http://storm.apache.org/

Apache Storm Official Website (2017). Storm v.1.1.0 documentation. Retrieved on 20 May 2017 from http://storm.apache.org/releases/current/Powered-By.html

Arora, D., Li, K. F., & Loffler, A. (2016, March). Big Data Analytics for Classification of Network Enabled Devices. In *Advanced Information Networking and Applications Workshops (WAINA), 2016 30th International Conference on* (pp. 708-713). IEEE.

Ashford, W. & johnson, T., J. (2016). Searching for Meaning: Advanced security tools are creating mounds of information, but how do infosec pros parse their meaning? With advanced analytics. *Search Security Techtarget*. Retrieved from http://searchsecurity.techtarget.com/

Ashton, K. (2009). That 'internet of things' thing. *RFiD Journal*, *22*(7), 97-114.

Bacon, M. (2017). St. Jude Medical finally patches vulnerable medical IoT devices. *Search Security Techtarget*. Retrieved from http://searchsecurity.techtarget.com/news/450410935/St-Jude-Medical-finally-patches-vulnerable-medical-IoT-devices

BBC News. (2015). Not in front of the telly: Warning over 'listening' TV. *BBC News/Technology*. Retrieved from http://www.bbc.com/news/technology-31296188

Bhatt, S., Manadhata, P. K., & Zomlot, L. (2014). The operational role of security information and event management systems. *IEEE Security & Privacy*, *12*(5), 35-41.

Blankenship, J. (2016). Market Overview: Security Analytics Platforms. *Forrester*. Retrieved from https://baydynamics.com/content/uploads/2016/05/Market_Overview_Security_Analytics_Platforms.pdf

Bloomberg (2017). Company Overview of iSight Security, Inc. *Bloomberg.* Retrieved on May 2017 from https://www.bloomberg.com/research/stocks/private/snapshot.asp?privcapId=102230684

Borum, R., Felker, J., Kern, S., Dennesen, K., & Feyes, T. (2015). Strategic cyber intelligence. Information & Computer Security, 23(3), 317-332.

Brewer, R. (2014). Advanced persistent threats: minimising the damage. *Network security*, *2014*(4), 5-9.

Bro Project (2017). TCP Scan detection. *Documentation of Bro 2.5*. TCP Scan detection. Retrieved on 21 May 2017 from https://www.bro.org/sphinx/scripts/policy/misc/scan.bro.html

Bussa, T., Litan, A. & Phillips, T. (2016). Market Guide for User and Entity Behavior Analytics. Retrieved from Gartner Research database (https://www.gartner.com/doc/3538217/market-guide-user-entity-behavior).

Bussa, T., Litan, A. & Ahlm, E. (2016). The Fast-Evolving State of Security Analytics, 2016. Retrieved from Gartner Research database (https://www.gartner.com/doc/3274217/fastevolving-state-security-analytics-).

Cardenas, A. A., Manadhata, P. K., & Rajan, S. P. (2013). Big data analytics for security. *IEEE Security & Privacy*, *11*(6), 74-76.

Casey, E., Blitz, A., & Steuart, C. (2014). Digital evidence and computer crime.

Cates, S. (2015). The evolution of security intelligence. *Network Security*, *2015*(3), 8-10.

Central Intelligence Agency. (1995). A Definition of Intelligence. *Central Intelligence Agency of USA*. Retrieved from https://www.cia.gov/library/center-for-the-study-of-intelligence/kent-csi/vol2no4/html/v02i4a08p_0001.htm

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, *36*(4), 1165-1188.

Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*, 314-347.

Cherdantseva, Y., & Hilton, J. (2013, September). A reference model of information assurance & security. In *Availability, reliability and security (ares), 2013 eighth international conference on* (pp. 546-555). IEEE.

Chiang, R. H., Goes, P., & Stohr, E. A. (2012). Business intelligence and analytics education, and program development: A unique opportunity for the information systems discipline. *ACM Transactions on Management Information Systems (TMIS)*, 3(3), 12.

Chuvakin, A. (2013) Network Forensics Defined? Gartner Blog Network. Retrieved on May 2017 from http://blogs.gartner.com/anton-chuvakin/2013/01/29/network-forensics-defined/.

Cloud Security Alliance. (2010). Expanded Top Ten Big Data Security and Privacy Challenges. *Cloud Security Alliance*. Retrieved on January 2017 from https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Expanded_Top_Ten_Big_Data_Security_and_Privacy_Challenges.pdf

Cloud Security Alliance. (2014). Big Data Taxonomy. *Cloud Security Alliance*. Retrieved on February 2017 from https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Taxonomy.pdf

Cloud Security Alliance. (2016). Big Data Security and Privacy Handbook 2016; 100 Best practices in Big Data Security and Privacy. *Cloud Security Alliance*. Retrieved on January 2017 from https://downloads.cloudsecurityalliance.org/assets/research/big-data/BigData_Security_and_Privacy_Handbook.pdf

Cloud Security Alliance. (2016). Big Data Analytics for Security Intelligence. *Cloud Security Alliance*. Retrieved on January 2017 from https://downloads.cloudsecurityalliance.org/initiatives/bdwg/Big_Data_Analytics_for_Security_Intelligence.pdf

Cohen, F. (1987). Computer viruses: theory and experiments. *Computers & security*, *6*(1), 22-35.

Darkreading. (2012). A case study in security big data analysis: at the RSA conference, Zions Bancorporation showed how Hadoop and BI analytics can power better security intelligence**.** Retrieved from http://www.darkreading.com/analytics/security-monitoring/a-case-study-in-security-big-data-analysis/d/d-id/1137299

Davenport, T. H. (2006). Competing on analytics. *Harvard business review*, *84*(1), 98.

Dean, J., & Ghemawat, S. (2008). MapReduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107-113.

Delta Consultants Company Official Website (2017). Blacklist IP Addresses Live Database Real-time. Retrieved on 17 May 2017 from https://myip.ms/files/blacklist/general/latest_blacklist.txt

Directive (EU) 2016/1148 of the European parliament and of the council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union [2016] OJ L194/1-30

Dog, S. E., Tweed, A., Rouse, L., Chu, B., Qi, D., Hu, Y., ... & Al-Shaer, E. (2016, August). Strategic Cyber Threat Intelligence Sharing: A Case Study of IDS Logs. In *Computer Communication and Networks (ICCCN), 2016 25th International Conference on* (pp. 1-6). IEEE.

Edwards, S., Ford, R., & Szappanos, G. (2015). Effectively Testing APT Defences. In *Proceedings of the 25th Virus Bulletin International Conference*.

Elmellas, J. (2016). Knowledge is power: the evolution of threat intelligence. *Computer Fraud & Security*, *2016*(7), 5-9.

European Union Agency for Network and Information Security - ENISA. (2016). Incentives and barriers for the cyber insurance market in Europe.. Retrieved from https://www.enisa.europa.eu/publications/cyber-insurance-recent-advances-good-practices-and-challenges/at_download/fullReport

Evans, B. & Graves, T.(2014). Spark and Storm at Yahoo! *Slideshare.net.* Retrieved on 10 May 2017 from https://www.slideshare.net/ChicagoHUG/yahoo-compares-storm-and-spark

Federal Trade Commission – FTC. (2017). IoT Home Inspector Challenge. Retrieved from https://www.ftc.gov/iot-home-inspector-challenge

Fetjah, L., Benzidane, K., El Alloussi, H., El Warrak, O., Jai-Andaloussi, S., & Sekkaki, A. (2016, June). Toward a Big Data Architecture for Security Events Analytic. In *Cyber Security and Cloud Computing (CSCloud), 2016 IEEE 3rd International Conference on* (pp. 190-197). IEEE.

Food and Drug Administration (FDA). (2016). *Postmarket Management of Cybersecurity in Medical Devices*. Retrieved on May 2017 from http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM482022.pdf.

Fortscale (2016). Advanced User Behavior Analytics. Official website of Fortscale enterprise. Retrieved from https://fortscale.com/solutions/

Frei, S. (2014). The Known Unknowns. *Update*.

Gantz, J., & Reinsel, D. (2011). Extracting value from chaos. *IDC iview*, *1142*(2011), 1-12.

Gartner, Inc. (2016). Gartner's 2016 Hype Cycle for Emerging Technologies Identifies Three Key Trends That Organizations Must Track to Gain Competitive Advantage. *Gartner Newsroom*. Retrieved on 25 May 2017 from http://www.gartner.com/newsroom/id/3412017

Ghemawat, S., Gobioff, H., & Leung, S. T. (2003, October). The Google file system. *ACM SIGOPS operating systems review* (Vol. 37, No. 5, pp. 29-43). ACM.

Goetz, P. T. (August 2014). Apache storm vs. Spark Streaming. *Slideshare.net.* Retrieved on 20 May 2017 from https://www.slideshare.net/ptgoetz/apache-storm-vs-spark-streaming

Google (2017). Frequently asked questions for Google Public DNS. *Google Public DNS*. Retrieved on 21 May 2017 from https://developers.google.com/speed/public-dns/faq#dnshttps_dnssec

Grahn, K., Westerlund, M., & Pulkkis, G. (2017). Analytics for Network Security: A Survey and Taxonomy. *In Information Fusion for Cyber-Security Analytics* (pp. 175-193). Springer International Publishing.

Gupta, A., Birkner, R., Canini, M., Feamster, N., Mac-Stoker, C., & Willinger, W. (2016). Network Monitoring as a Streaming Analytics Problem. In *Proceedings of the 15th ACM Workshop on Hot Topics in Networks-HotNets' 16*. Association for Computing Machinery (ACM).

Harding, L. (2017). What we know about Russia's interference in the US election. *Theguardian website.* Retrieved on May 2017 from https://www.theguardian.com/us-news/2016/dec/16/qa-russian-hackers-vladimir-putin-donald-trump-us-presidential-election

Hardof T. (2016). Making Web Application Security Statistics Meaningful. *Whitehat Security blogs.* Retrieved on May 2017 from https://www.whitehatsec.com/blog/web-application-security-stats-report-2016/

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98-115.

Herschel, G., Linden, A., & Kart, L. (2015). Magic quadrant for advanced analytics platforms. *Gartner Report G*, *270612*. Retrieved from Gartner Research database (http://www.gartner.com)

Hewlett Packard Enterprise – HPE (2017). HPE Acquires Niara to Enhance Security at the Intelligent Edge. *HPE Newsroom*. Retrieved on May 2017 from https://news.hpe.com/hpe-acquires-niara-to-enhance-security-at-the-intelligent-edge/

Hilton, S. (2016). Dyn Analysis Summary Of Friday October 21 Attack. *Dyn*. Retrieved from http://dyn.com/blog/dyn-analysis-summary-of-friday-october-21-attack

IKANOW. (2015). Cyber threat analytics versus threat intelligence. *IKANOW Editorial*. Retrieved from https://ikanow.com/cyber-threat-analytics-versus-threat-intelligence

International Business Machines Corp. - IBM (2016). IT executive guide to Security Intelligence: Transitioning from log management and SIEM to state-of-the-art Security with advanced IBM Sense Analytics Engine. *International Business Machines Corp.* Retrieved from https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WGW03268USEN

Internet Live Stats (2017). Internet Live Stats. *Real Time Statistics Project.* Retrieved on February and May 2017 from http://www.internetlivestats.com/one-second/

Janessa Rivera Gartner Rob van der Meulen (2015, April 16). *Gartner Says Security Analytics May Be Key in Breach Detection* (ID: 3030818). Retrieved from Gartner database.

Johnson, T., J. (2015). User behavioral analytics tools can thwart security attacks. *Search Security Techtarget.* Retrieved from http://searchsecurity.techtarget.com/feature/User-behavioral-analytics-tools-can-thwart-security-attacks

Kissel, R. (2013). Glossary of key information security terms. NIST Interagency Reports NIST IR, 7298(3)

Khan, A. N., Kiah, M. M., Ali, M., Madani, S. A., & Shamshirband, S. (2014). BSS: block-based sharing scheme for secure data storage services in mobile cloud environment. *The Journal of Supercomputing*, *70*(2), 946-976.

Krebs, B. (2016). KrebsOnSecurity Hit With Record DDoS. Retrieved from https://krebsonsecurity.com/2016/09/krebsonsecurity-hit-with-record-ddos

Lafuente, G. (2015). The big data security challenge. *Network security*, *2015*(1), 12-14.

Lan, L., & Jun, L. (2013, December). Some special issues of network security monitoring on big data environments. In *Dependable, Autonomic and Secure Computing (DASC), 2013 IEEE 11th International Conference on* (pp. 10-15). IEEE.

Leibiusky, J., Eisbruch, G., & Simonassi, D. (2012). *Getting started with storm*. " O'Reilly Media, Inc."

Lemos, R. (2015). Are SIEM systems delivering on advanced analytics? *Search Security Techtarget.* Retrieved from http://searchsecurity.techtarget.com/video/Are-SIEM-systems-delivering-on-advanced-analytics

Lemos, R. (2015). SIEM systems: Using analytics to reduce false positives: Combining data from a variety of sources with better analytics can reduce workloads. *Search Security Techtarget.* Retrieved from http://searchsecurity.techtarget.com/tip/SIEM-systems-Using-analytics-to-reduce-false-positives

Lim, E. P., Chen, H., & Chen, G. (2013). Business intelligence and analytics: Research directions. *ACM Transactions on Management Information Systems (TMIS)*, *3*(4), 17.

Litan, A. & Nicolett, M. (2014). *Market Guide for User Behavior Analytics*. Retrieved from Gartner Research database (http://www.gartner.com).

Liu, H. (2013). Big data drives cloud adoption in enterprise. *IEEE internet computing*, *17*(4), 68-71

Lobato, A., Lopez, M. A., & Duarte, O. C. M. B. (2016). An accurate threat detection system through real-time stream processing. *Grupo de Teleinformática e Automaçao (GTA), Universidade Federal do Rio de Janeiro (UFRJ), Tech. Rep. GTA-16-08*.

LogRhythm. (2015). LogRhythm Accelerates Detection and Response to Cyber Threats with New Case Management and Advanced Search Features. *LogRhythm*. Retrieved from https://logrhythm.com/about/press-releases/logrhythms-new-case-management/

Mahalanobis, P. C. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, *2*, 49-55.

Mahmood, T., & Afzal, U. (2013, December). Security analytics: Big data analytics for cybersecurity: A review of trends, techniques and tools. In *Information assurance (ncia), 2013 2nd national conference on* (pp. 129-134). IEEE.

Markets and Markets (2016). Multifactor Authentication Market worth 12.51 Billion USD by 2022. Markets and Markets. Retrieved on May 2017 from http://www.marketsandmarkets.com/PressReleases/multi-factor-authentication.asp

Marz, N., & Warren, J. (2015). *Big Data: Principles and best practices of scalable realtime data systems*. Manning Publications Co..

Mattern, T., Felker, J., Borum, R., & Bamford, G. (2014). Operational levels of cyber intelligence. *International Journal of Intelligence and CounterIntelligence*, *27*(4), 702-719.

Mavridis, I. (2015). Internet Information Security. Hellenic Academic Libraries Link.

Mell, P., & Grance, T. (2011). The NIST definition of cloud computing.

Miloslavskaya, N., Senatorov, M., Tolstoy, A., & Zapechnikov, S. (2014). Information Security Maintenance Issues for Big Security-Related Data. In *Future Internet of Things and Cloud (FiCloud), 2014 International Conference on* (pp. 361-366). IEEE.

Naydenov, R., Liveri, D., Dupre, L., Chalvatzi, E. & Skouloudi, C. (2015). Big Data Security - Good Practices and Recommendations on the Security of Big Data Systems. *European Union Agency for Network and Information Security - ENISA*. Retrieved from https://www.enisa.europa.eu/publications/big-data-security/at_download/fullReport

Newman, L., H. (2014). Pretty Much Every Smart Home Device You Can Think of Has Been Hacked. *Slate Future Tense Blog.* http://www.slate.com/blogs/future tense/2014/12/30/the internet of things is a long way from being secure.html

Open Threat Exchange API (2017). *Alien Vault*. Retrieved on May 2017 from https://otx.alienvault.com/

Palmer, G. (2001). A road map for digital forensic research. *In First Digital Forensic Research Workshop, Utica, New York* (pp. 27-30).

Pandey, S., & Nepal, S. (2013). Cloud computing and scientific applications—big data, scalable analytics, and beyond. *Future Generation Computer Systems*, *7*(29), 1774-1776

Pescatore, J., & Young, G. (2009). Defining the next-generation firewall. *Gartner RAS Core Research Note*.

Petersen, C. (2015). Surfacing Critical Cyber Threats Through Security Intelligence. LogRhythm. Retrieved from https://logrhythm.com/pdfs/whitepapers/lr-security-intelligence-maturity-model-ciso-whitepaper.pdf

Ponemon Institute LLC. (2015). The Cost of Malware Containment. *Ponemon Institute Research Report.* Retrieved from http://www.ponemon.org/local/upload/file/Damballa%20Malware%20Containment%20FINAL%203.pdf

PricewaterhouseCoopers - PwC. (2016). Moving forward with cybersecurity and privacy: How organizations are adopting innovative safeguards to manage threats and achieve competitive advantages in a digital era. *PriceWaterhouseCoopers*. Retrieved from https://www.pwc.ch/en/publications/2016/moving_forward_with_cybersecurity_and_privacy_swiss_web.pdf

PricewaterhouseCoopers - PwC (2017). Global Economic Crime Survey for 2016. *PriceWaterhouseCoopers.* Retrieved on May 2017 from https://www.pwc.com/gx/en/economic-crime-survey/pdf/GlobalEconomicCrimeSurvey2016.pdf

PricewaterhouseCoopers - PwC. (2016). Toward new possibilities in threat management: How businesses are embracing a modern approach to threat management and information sharing. *PriceWaterhouseCoopers*. Retrieved from

https://www.pwc.com/gx/en/issues/cyber-security/information-security-survey/assets/gsiss-report-cybersecurity-privacy-possibilities.pdf

RapidMiner**.** (2014). An Introduction to Advanced Analytics. *RapidMiner***.** Retrieved from https://rapidminer.com/wp-content/uploads/2014/04/advanced-analytics-introduction.pdf

Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1-88

Rivera, J. & Meulen, R. (2015). Gartner Says Security Analytics May Be Key in Breach Detection. Retrieved from Gartner Research database (http://www.gartner.com).

Rouse, M., Martinek, L. & Stedman, C. (2014). Big data analytics. *Search Business Analytics Techtarget*. Retrieved from http://searchbusinessanalytics.techtarget.com/definition/big-data-analytics

Rouse, M., & Wigmore, I. (2015). Security intelligence (SI). *WhatIs Techtarget*. Retrieved from http://whatis.techtarget.com/definition/security-intelligence-SI

Rouse, M., & Wigmore, I. (2015). Threat intelligence (Cyber Threat Intelligence). *WhatIs Techtarget.* Retrieved from http://whatis.techtarget.com/definition/threat-intelligence-cyber-threat-intelligence

Saltzer, J. H., & Schroeder, M. D. (1975). The protection of information in computer systems. *Proceedings of the IEEE*, *63*(9), 1278-1308.

Scarfone, K. (2016). Say Hello, Again, to SIEM. The technological capabilities of security information and event management tools have evolved. Get up to date on what SIEM systems can do for you now. *Search Security Techtarget.* Retrieved from http://searchsecurity.techtarget.com/ehandbook/SIEM-products-and-capabilities-you-need-now

Scarfone, K. (2016). Introduction to SIEM services and products. *Search Security Techtarget*. Retrieved on 10 May 2017 from http://searchsecurity.techtarget.com/feature/Introduction-to-SIEM-services-and-products

Schurgot, M. R., Shinberg, D. A., & Greenwald, L. G. (2015, June). Experiments with security and privacy in IoT networks. In *World of Wireless, Mobile and Multimedia Networks (WoWMoM), 2015 IEEE 16th International Symposium on a* (pp. 1-6). IEEE.

Schwartz, S. (2016). Managing Medical Device Cybersecurity in the Postmarket: At the Crossroads of Cyber-safety and Advancing Technology [Blog post]. Retrieved from http://blogs.fda.gov/fdavoice/index.php/2016/12/managing-medical-device-cybersecurity-in-the-postmarket-at-the-crossroads-of-cyber-safety-and-advancing-technology/

Seo, K., & Kent, S. (2005). Security architecture for the internet protocol. Retrieved on May 2017 from https://tools.ietf.org/html/rfc4301

Shashanka, M., Shen, M. Y., & Wang, J. (2016, December). User and entity behavior analytics for enterprise security. *In Big Data (Big Data), 2016 IEEE International Conference on* (pp. 1867-1874). IEEE.

Shirey, R. W. (2007). Internet security glossary, version 2. Retrieved on May 2017 from https://tools.ietf.org/html/rfc4949

Shirey, R. (2003). RFC 2828–Internet security glossary, 2000. Retrieved on May 2017 from https://www.ietf.org/rfc/rfc2828.txt

Shubham S. (2016). HADOOP ECOSYSTEM. *Edureka blogs for Big Data Analytics*. Retrieved on 15 May 2017 from https://www.edureka.co/blog/hadoop-ecosystem

Simson, G. (2017). Network Forensics: Tapping the Internet. Retrieved in May 2017 from http://www.oreillynet.com/pub/a/network/2002/04/26/nettap.html

Statistic Brain Website (2017). Statistic Brain. *Statistic Brain Research Institute.* Retrieved on February and May 2017 from http://www.statisticbrain.com/

Stergiou, C., & Psannis, K. E. (2017). Recent advances delivered by mobile cloud computing and internet of things for big data applications: a survey. *International Journal of Network Management*, *27*(3).

Stergiou C.& Psannis K. (2016). "Recent advances delivered by Mobile Cloud Computing and Internet of Things for Big Data applications: a survey," International Journal of Network Management, pp. 1-12.

Stevanovic, M., Pedersen, J. M., D'Alconzo, A., & Ruehrup, S. (2016). A method for identifying compromised clients based on DNS traffic analysis. *International Journal of Information Security*, 1-18.

Terzi, D. S., Terzi, R., & Sagiroglu, S. (2015). A survey on security and privacy issues in big data. In *Internet Technology and Secured Transactions (ICITST), 2015 10th International Conference for* (pp. 202-207). IEEE.

Tirosh, A. & Proctor, E., P. (2016). Shift Cybersecurity Investment to Detection and Response. Retrieved from Gartner Research database (https://www.gartner.com/doc/3183622/shift-cybersecurity-investment-detection-response).

Thomas P. (2015). Why big data can help to keep planes in the air. *Blogs of **SAP** (Systemanalyse und Programmentwicklung).* Retrieved on 15 May 2017 from https://blogs.sap.com/2015/01/20/why-airlines-need-to-keep-planes-in-the-air/

Thuraisingham, B. (2015, March). Big data security and privacy. In *Proceedings of the 5th ACM Conference on Data and Application Security and Privacy* (pp. 279-280). ACM.

Toshniwal, A., Taneja, S., Shukla, A., Ramasamy, K., Patel, J. M., Kulkarni, S., ... & Bhagat, N. (2014, June). Storm@ twitter. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 147-156). ACM.

Turban, E., Sharda, R., Aronson, J. E., & King, D. (2008). *Business intelligence: A managerial approach*. New Jersey: Pearson Prentice Hall.

Tutorialspoint.com (2017). A tutorial of Apache Storm. *Tutorialpoint.com*. Retrieved on May 2017 from https://www.tutorialspoint.com/apache_storm/index.htm

Ussath, M., Jaeger, D., Cheng, F., & Meinel, C. (2016). Pushing the Limits of Cyber Threat Intelligence: Extending STIX to Support Complex Patterns. In *Information Technology: New Generations* (pp. 213-225). Springer International Publishing.

Van Kranenburg, R. (2008). *The Internet of Things: A critique of ambient technology and the all-seeing network of RFID*. Institute of Network Cultures.

Vaughan, J. (2016). Hadoop, Kafka creators big on big data streaming analytics. *TechTarget podcasts*. Retrived on May 2017 from http://searchdatamanagement.techtarget.com/podcast/Hadoop-Kafka-creators-big-on-big-data-streaming-analytics

Verisign (2015). Introducing Verisign Public DNS: A Free Recursive DNS Service That Respects Your Privacy. Retrieved on 21 May 2017 from http://www.circleid.com/posts/20150929_verisign_public_dns_free_dns_service_respects_privacy/

Verizon. (2016). 2016 Data Breach Investigations Report. Retrieved on 1 June 2017 from http://www.verizonenterprise.com/resources/reports/ rp_DBIR_2016_Report_en_xg.pdf

Virvilis, N., Serrano, O., & Dandurand, L. (2014). Big Data analytics for sophisticated attack detection. *Isaca Journal*, *3*, 22-25.

Watson, H. J., & Wixom, B. H. (2007). The current state of business intelligence. *Computer*, *40*(9).

White, T. (2012). *Hadoop: The definitive guide*. " O'Reilly Media, Inc.".

Yoon, S., Park, H., & Yoo, H. S. (2015). Security issues on smarthome in IOT environment. In *Computer Science and its Applications* (pp. 691-696). Springer Berlin Heidelberg.

Zikopoulos, P., Parasuraman, K., Deutsch, T., Giles, J., & Corrigan, D. (2012). *Harness the power of big data The IBM big data platform*. McGraw Hill Professional.

Zubair, A. (2016). More Security Predictions: 2017 Will be a Mixed Bag. *Whitehatsecurity*. Retrieved from https://www.whitehatsec.com/blog/2017-will-be-a-mixed-bag/