DATA GROWTH AND ITS IMPACT ON CYBERSECURITY AND ARTIFICIAL INTELLIGENCE

Cristian CUCU[1], Gheorghe GAVRILOAIA[2]
[1,2]University of Pitesti, Romania
[1]ccris73@gmail.com, [2]ggavriloaia@gmail.com

Keywords: data growth, cybersecurity, Moore's law, artificial intelligence, digital transformation

Abstract: *Digitization is becoming the leading growth vector worldwide and is being incorporated into all economic sectors, with predictions that it will reach half of the world economy by 2021. The explosion of data started almost two decades ago and in such a short time the amount of collected data has created a new sector by itself: Data Economy. Such a fast development has led to a quest for identifying performant Artificial Intelligence methods to classify and analyze large volumes and offer the right information at the right time. The same data explosion has generated a huge demand for cybersecurity and data protection as core systems become online and the transaction data becomes more sensitive.*

1. INTRODUCTION

Digital transformation is demanding more and more complex IT systems and infrastrucures, becoming slowly but surely the main engine for the world economy with a very wide reach into all sectors: business, industry, education, health, military, agriculture, banking and financial markets, insurance, oil and gas, public administration, energy and utilities, etc.

One of the most discussed topic in every business entity is now „digital transformation" and prestigious IDC researchers have predicted for 2018 that 2/3 of Global 2000 CEOs will focus their strategies on Digital Transformation [1] while by 2020 they will spend over 50% of their budgets for digital transformation initiatives [2] to follow that strategy, as shown in Figure 1.

This new wave is backed by a massive investment of almost 6 trillion USD in only three years, from 2018-2021 [3] with an expected result of digitizing half of the world economy by 2021 [4]. The investments also create a multiplied innovation environment that accelerate the trend of both venture investments and technology advancements towards a faster adoption of digitization into the business processes.

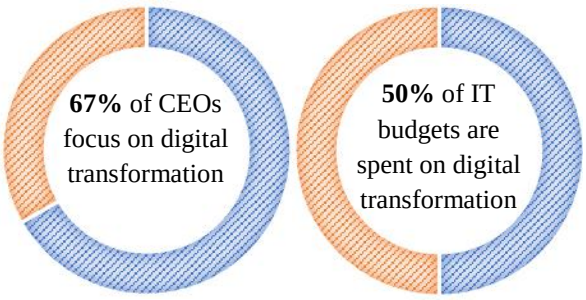

Fig. 1 – Digital transformation efforts [1,2]

In addition to such a massive reach in such a short time for sectors that have not been traditionally favoring digitization, the trend has created a new digital sector which is growing very powerful on its own, to support the demand in all these sectors.

2. DATA EXPLOSION

Such a massive evolution of digitization for current business models and processes is generating significant data from every node of that process, starting from authorization all the way to logging and auditing all transactions.

This massive increase in generated data by such complex IT systems needs to be handled by similarly powerful storage and support systems. The problem of data storage has lead to different types of research and the recent human genome revolution also touched this subject by identifying that one gram of DNA can theoretically hold 455 exabytes and it could last for 2,000 years at 10^0C or even 2 million years at -18^{0}C [5].

A simple yet representative problem is that we need to define new words for data volume beyond the commonly used gigabyte or terabyte like exabyte (10^{18} bytes), zettabyte (10^{21} bytes) or yottabyte (10^{24} bytes).

The world's capacity to store 2,6 exabytes (EB) was first reached in 1986 [6] and this was the equivalent of one CD per person on Earth. From the beginning of mankind up until 2003, only 5 EB of information were generated while 10 years later, the same amount is being generated in two days [7]. The Figure 2 shows the evolution of data creation and the split between digital and analog data.

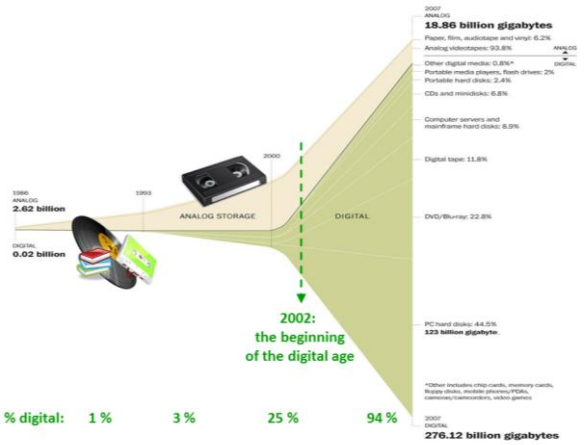

Fig. 2 – Digital age storage, adaptation [8]

We are now in the zettabyte (ZB) era based on measuring the World's stored capacity of digital information in 2012 of 1 ZB [9] and the global IP traffic that exceeded 1,2 ZB in 2016. The same source has predicted that the data explosion will follow Moore's law, doubling every two years but later it was proven that data is growing faster than processing power development.

Internet associated energy consumption has reached 2% in 2015 [10] and the global emissions generated by all IT activities are estimated to reach 12% by 2020 [11].

Usually the positive scale is measured and recorded before negative measurements can be reached but in respect to yotta, the first measurement is that of yoctonewtons or 10^{-24} for atomic forces measured by the quantum physicists from NIST using trapped beryllium ions [12]. A similar value on the positive scale is expected to represent YottaBytes (YB) by 2030 as forecasted by Science Magazine.

In 2019 a new proposal has been formed to the International Bureau of Weights and Measures (BIPM) in Paris to have names approved for the upcoming need to define new values for storage: ronna and quecca that represent 10^{27} and respectively 10^{30}. The Bureau meets every three years and it is expected to have this topic under debate for the 2022 meeting. This would be the first addition form the last series approved in 1991 that has been referenced up to yotta. If these will be approved, we will be able to talk about RottaByte (RB) and QueccaByte (QB).

All this data turned into a market of its own, named "Data Economy" with rules similar to other industries but one main difference: the volume of the digitized information captures all domains and increases much faster than any other sector by itself.

No matter the originating source of data, it is being treated by all economic sectors as an emerging asset, treated similarly to other corporate assets such as equipment, buildings, inventory etc. The more things become interconnected, the more data they will generate and this will require more analysis.

The "Data Economy" is being defined as the global system where data is being collected, organized and transaction between specialized actors with the objective of monetizing the value resulted from the gathered information.

The increase in data is beyond Moore's law for processing power as presented in Figure 3 below, at an approximate pace of 64% per year having domains that store at even a higher pace such as DNA Sequencing (in the Exabyte area, increasing fast as genomes are being decoded faster) and particle acceleration data (CERN's Large Hadron Collider generates 1 Petabyte/sec [13,14]). Lower costs, approaching 1,000 USD for Genome sequencing is another factor that leads to more and more requests for genome decoding.

3. DATA GROWTH AND ARTIFICIAL INTELLIGENCE

As mentioned above, the more interconnected devices, the more data is being generated and more analysis is needed. Data scientists have always worked on data samples and the sudden wealth of big data has allowed for a shift to work on real datasets. The new trend has shifted again, when Artificial Intelligence (AI) started to play a big role by allowing sampling that is becoming more and more relevant in addition to the regular transformation of data with bigger blocks and less errors.

This evolution of data over hardware development in the past couple of years has accelerated in both ways: data is growing even faster while hardware is slowing down, slowly underperforming in Moore's law as shown in Figure 3 below.

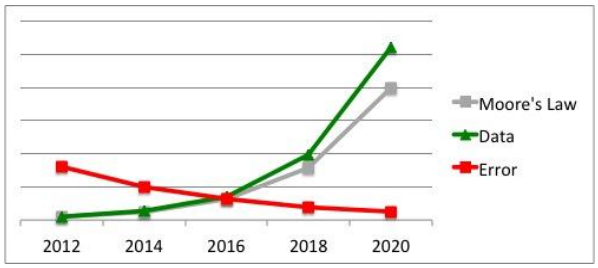

Fig. 3 – Data growth versus Moore's law [14]

This is mainly due to the mainstream approach to Big Data and AI concepts and implementations which put extra demand on hardware platforms which in turn have very little room to become technologically denser and more efficient.

4. DATA GROWTH AND THE DEMAND FOR CYBERSECURITY

Together with the increase in data, there is an even more significant increase in risk.

Accelerating the pace at which data becomes available without having a strong set of rules that are generally applicable generates a risk exposure that is being measured in several cybersecurity reports. As we are heading towards 200 billion connected devices in 2020 as predicted by Intel [15] the amount of vulnerable data is being demonstrated by the number of stolen digital records which in the first half of 2017 was 1.9 billion, or approximately 100 per second [15]. This figure has rounded up to 2.5 billion for the entire 2017. The same threat counted almost 3.5 billion records stolen in the first 6 months of 2018 approximately 70% more than H1 2017 as shown in the Gemalto Breach Level Index [16]. The same index shows a total of close to 15 billion records stolen since 2013, which is around 2 records per person worldwide, stolen already. In 2019, „Collection #1" has become publicly available, consisting of 2,7 billion records exposing almost 772 million email addresses and 21 million passwords [17]. Apparently, the same hacker claims to have 6 more batches of data [17].

The sensitivity of transaction data increases as more and more industries are digitizing their core operations, which used more secure channels in the past. The need for always-on data is forcing companies to work online, exposing to unknown risks data that previously has been somehow safely kept inside isolated corporate networks. These companies, to manage the exposed data, deploy more and more cybersecurity technologies to protect against the vulnerabilities and increased risk that comes with the new type of presence. In order to keep track of all events and understand the cyber risks, they hire dedicated personnel to manage these devices and implement a cybersecurity event management solution (SIEM) to maintain a dashboard look at day to day threats. The more data is being exposed, the more security measures are being implemented and more hardware is being deployed to support the SIEM and its ecosystem to capture all events in order to identify potential attacks. This spiral effect results in unmanageable amount of cybersecurity data that rarely remains relevant in an always-on scenario. Security devices, including SIEMs generate false-positives that add more stress to the analysts that have to make sense of all the information that is being analyzed [15].

Cybersecurity provoked damages are expected to increase from $3 trillion in 2015 to $6 trillion in 2019 while the job market is showing 0% unemployment rate since before 2014 and the demand for professionals has increased from 1 million openings in 2015 to an estimated of 6 million job openings in 2019 [18] based on Palo Alto research Center.

These trends are here to stay as not only data is exploding in the new environment but also the data consumers are growing at a similar pace: 2 billion internet users in 2015, 4 billion in 2018 and the predictions from Cybersecurity Ventures show an expected 6 billion users by 2022 or 75% of the population and 7.5 billion in 2030, a staggering 90% of the projected world population 8.5 billion of six or more years old in that year [19].

The alignment of powerful trends like the increasing demand in specialized workforce and unemployment, the strong growth in generated data and data consumers, the increase in cybersecurity incidents is accelerating the request for the use of automated tools to compensate the lack of risk management in the cyber domain. Such tools may employ rules to categorize the data and use AI algorithms to aggregate information, generate rules and run them automatically in order to self-adapt the protection ecosystem.

5. CONCLUSIONS

The birth of Data Economy has been well received by our traditional economy and the increase in digital transformation projects is showing us that this trend not only is strong now but is increasing at a very high pace, beyond original predictions as it captures all legacy economic sectors.

Digital transformation is more about transformation that it is about digital. All relevant studies have shown that current organizational cultures are averse to change and to adopt new business processes and 70% of all digital transformation projects have failed in reaching their objective [20] because of this very reason.

All organizations will have to adapt their traditional processes and make room for digitization that will bring more efficiency in their operations if implemented properly.

Together with this digitization process, cybersecurity becomes more relevant and newer and faster technologies will have to be deployed in conjunction with AI in order to make sense of all the rising threats and new vulnerabilities within the IT systems that start to control every aspect of every organization.

6. REFERENCES

[1]. Business Wire Magazine, accessed 24.07.2019: https://www.businesswire.com/news/home/20151104006575/en/IDC-Reveals-Worldwide-Digital-Transformation-Predictions-Kicks

[2] Gil Press, Forbes, accessed 24.07.2019: https://www.forbes.com/sites/gilpress/2015/12/06/6-predictions-about-the-future-of-digital-transformation/#18370cac1102

[3] Sarah M., accessed 24.07.2019: www.idc.com/getdoc.jsp?containerId=prUS44430918

[4] IDC, accessed 24.07.2019: https://www.idc.com/events/futurescape

[5] Aron, Jacob, "Glassed-in DNA makes the ultimate time capsule". New Scientist, 2019

[6] Hilbert, M; López, P "The World's Technological Capacity to Store, Communicate, and Compute Information". *Science*. 332 (6025): 60–65

[7] Vance, Jeff. "Big Data Analytics Overview". Datamation. Retrieved 07.2019

[8] Washington Post, citing Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, *332*(6025), 60 –65.

[9] Gantz, John; Reinsel, David, accessed 2019: "The Digital Universe 2020: Big Data, Bigger Digital Shadows and Biggest Grow in the Far East".

[10] Newman, Kim, "ChipScaleReview", 2019

[11] Rong, H; Zhang, H; Xiao, S; Li, C; Hu, C, "Optimizing energy consumption for datacenters" Renewable & Sustainable Energy Reviews, 2016.

[12] M.J.Biercuk, H.Uys, J.W.Britton, A.P. VanDevender, J.J.Bollinger, „Ultrasensitive force and displacement detection using trapped ions", Nature Nanotechnology, 2010

[13] Melissa Gaillard, accessed 2018, https://home.cern/news/news/computing/cern-data-centre-passes-200-petabyte-milestone

[14] Ion Stoica, accessed 07.2019: https://amplab.cs.berkeley.edu/for-big-data-moores-law-means-better-decisions/

[15] John Bemis, accessed 07.2019: https://bmarkits.com/cybersecurity-big-data/

[16] https://breachlevelindex.com/, accessed 07.2019

[17] Victoria Song, https://gizmodo.com/mother-of-all-breaches-exposes-773-million-emails-21-m-1831833456, 2019

[18] Steve Morgan, accessed 07.2019: https://cybersecurityventures.com/cybersecurity-unemployment-rate/

[19] Steve Morgan, accessed 07.2019: https://cybersecurityventures.com/cybercrime-damages-6-trillion-by-2021/

[20] Behnam Tabrizi, Ed Lam, Kirk Girard, Vernon Irvin, https://hbr.org/2019/03/digital-transformation-is-not-about-technology, 2019, Harvard Business Review